

# BioCompute Objects on the High-performance Integrated Virtual Environment

---

JONATHON KEENEY, PH.D

ASSISTANT RESEARCH PROFESSOR, GEORGE WASHINGTON UNIVERSITY



**BioCompute**  
Objects

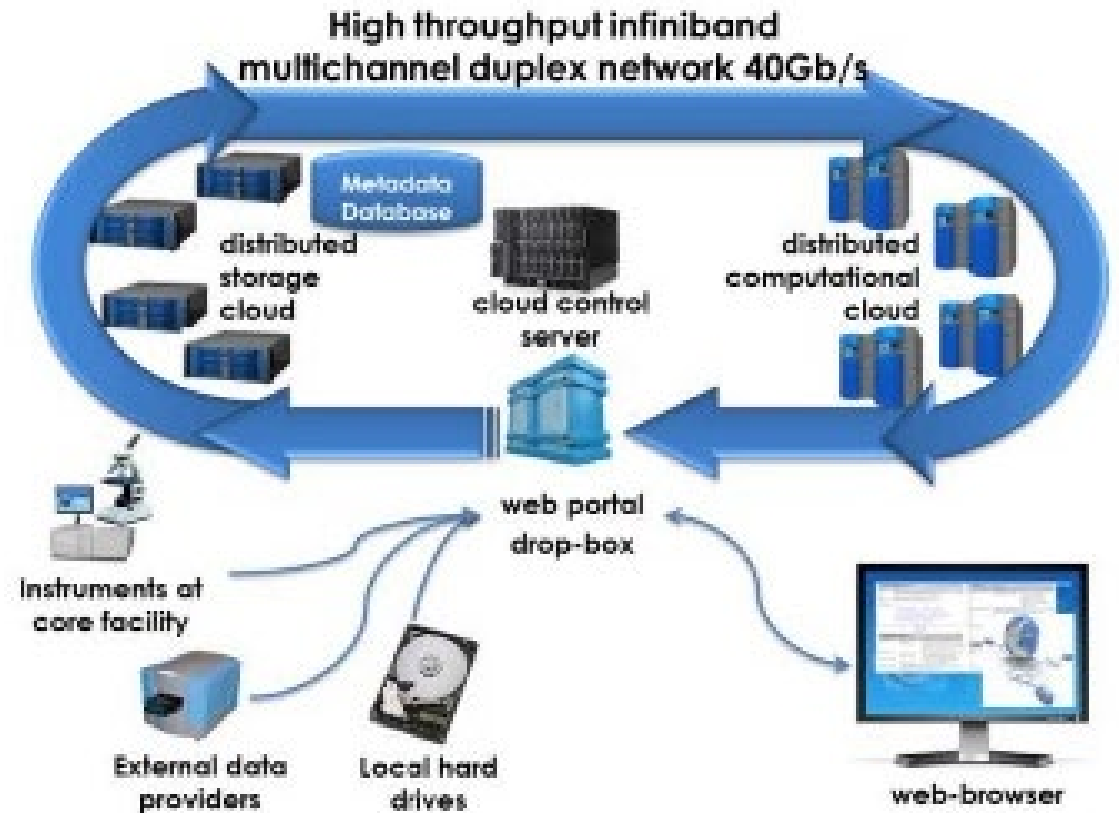
**Onc**  **MX**

**GlyGen** 

**Other projects:**  
Microbiome  
Neurogenomics

# High-performance Integrated Virtual Environment (HIVE)

- Platform for bioinformatic analysis of very big data
  - GUI wrapper for command line scripts
- Highly parallelizable
  - Storage and compute are distributed to many powerful servers
- Developed by Raja Mazumder and Vahan Simonyan
  - Code given to FDA
  - Now “Mission Critical” at FDA
  - GW maintains only publicly available version



# HIVE Infrastructure

---

## Interoperable

- Full featured API
- “HIVE Packs” can transfer computations and associated data between architectures
- Isolated environments for development, production, and scaling

## Secure

- Gauntlet of private sector tests
- Used by academic collaborators, FDA, and private sector collaborators

## Fast

- Tools designed and optimized for HIVE architecture
- Internet2 for data transfer



home help

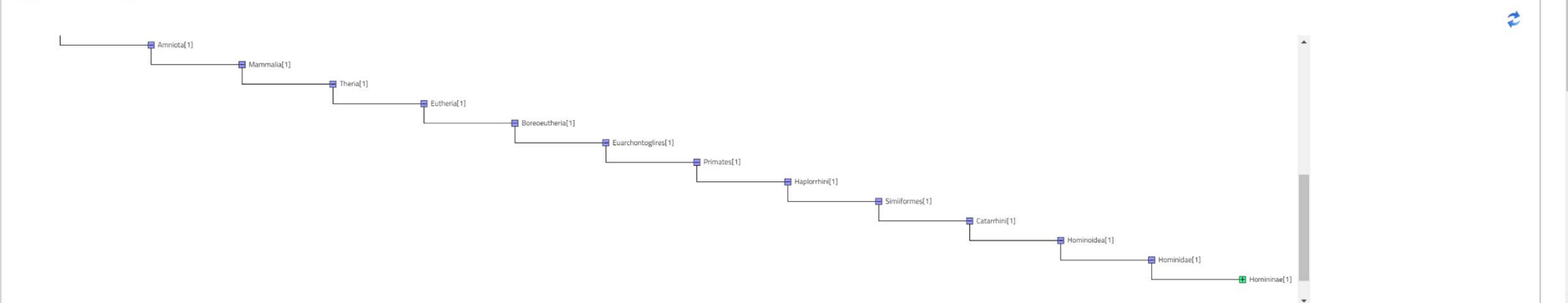
- Cut Edit
- HIVE Space
- All objects
- Inbox
  - genome\_assemblies\_genome\_fasta.tai
  - HCB
  - HumanGutDB\_Epilepsy.zip
  - IgG-H1b.zip
  - MicrobiomeData
  - ProteinsBiomarkersClass
- Trash
  - sample

all[90] folders[6] genomes[2] reads[5] annotations[1] files[37] computations[47] data-loading[32]

Convert Export Details

1-50 of 98 50

ID	Summary	Created
5466	<b>Glymps for glyco-proteomics 100%</b>	4/6/2020
5355	IgG-H1b.zip <b>Folder</b>	4/6/2020
5353	glymps_protease_library_version-1.0.0.0.txt	4/6/2020
5352	glymps_mass_library_version-1.0.0.0.txt	4/6/2020
5351	glymps_mod_library_version-1.0.0.0.txt	4/6/2020
5350	IgG-H1b.zip <b>File Upload 100%</b>	4/6/2020
5349	glymps_mass_library_version-1.0.0.0.txt, glymps_mod_library_version-1.0.0.0.txt, glymps_protease_library_version-1.0.0.0.txt <b>File Upload 100%</b>	4/6/2020
5344	<b>GlycoPeptideSearch 100%</b>	4/5/2020
5343	iggs.fasta	4/4/2020
5342	iggs.fasta <b>File Upload 100%</b>	4/4/2020
4810	<b>MS-GF+ 100%</b>	3/28/2020
4809	Example_Glycomics1 <b>MS-GF+ 100%</b>	3/28/2020
4808	<b>MS-GF+ 100%</b>	3/27/2020
4786	ProteinsBiomarkersClass <b>Folder</b>	3/27/2020
4785	'NM_004304.3.gb <b>ionAnnot type(s)</b>	3/26/2020
4784	NM_004304.3.gb 1 <b>Genomic sequence(s)</b>	3/26/2020
4783	genbank: NM_004304.3 <b>Downloader Engine 100%</b>	3/26/2020
4734	MicrobiomeData <b>Folder</b>	3/22/2020
4733	11566.S22.fastq 166468 <b>Nucleotide read(s)</b>	3/22/2020
4732	Experimental <b>Downloader Engine 81%</b>	3/22/2020
4731	11566.S21.fastq 172731 <b>Nucleotide read(s)</b>	3/22/2020



# HIVE Talks

---

## **Novel Approach for Identification of Defective Viral Genomes using NGS Data**

- Defective viral genomes (DVGs) are spontaneously generated by most viruses
- Some DVGs recognized as important triggers of antiviral innate immunity
- DVGs have implications for vaccine immunogenicity
- New HIVE tool: DVG-profiler
- Tool presentation, along with application in viral quasispecies analysis

## **Next talk: One-click RNAseq analysis**

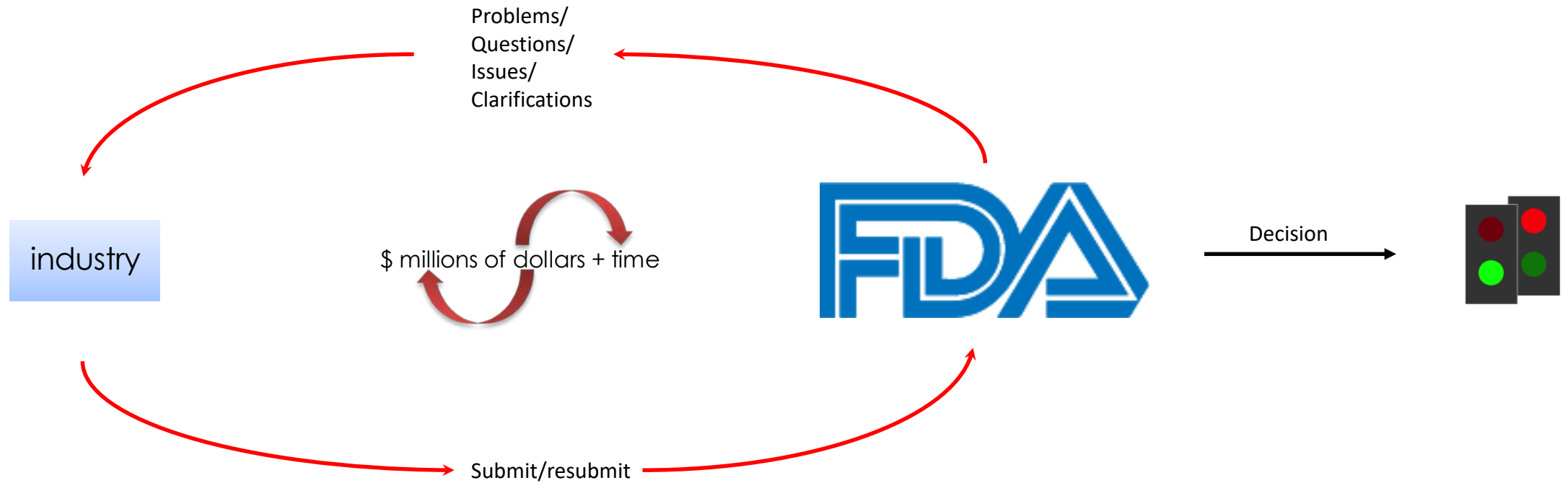
- Luis Santana-Quintero
- Date TBD



Konstantinos Karagiannis, Ph.D.

# Wasted Time and Money

---



# Solution: BioCompute

Experimental Design

Analysis Steps

Parameters

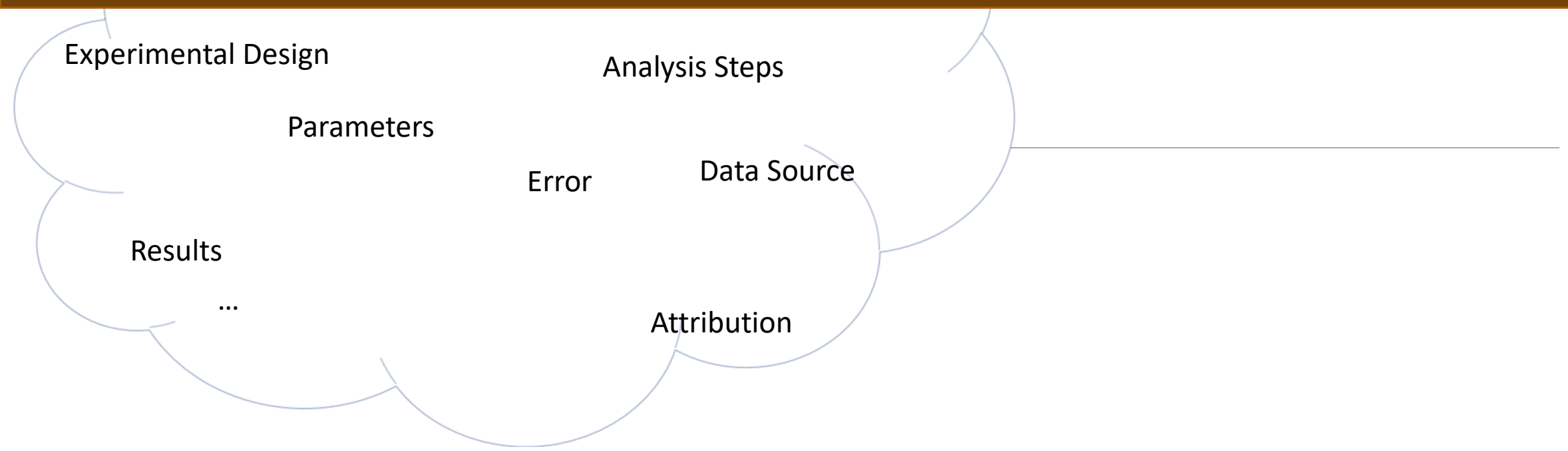
Error

Data Source

Results

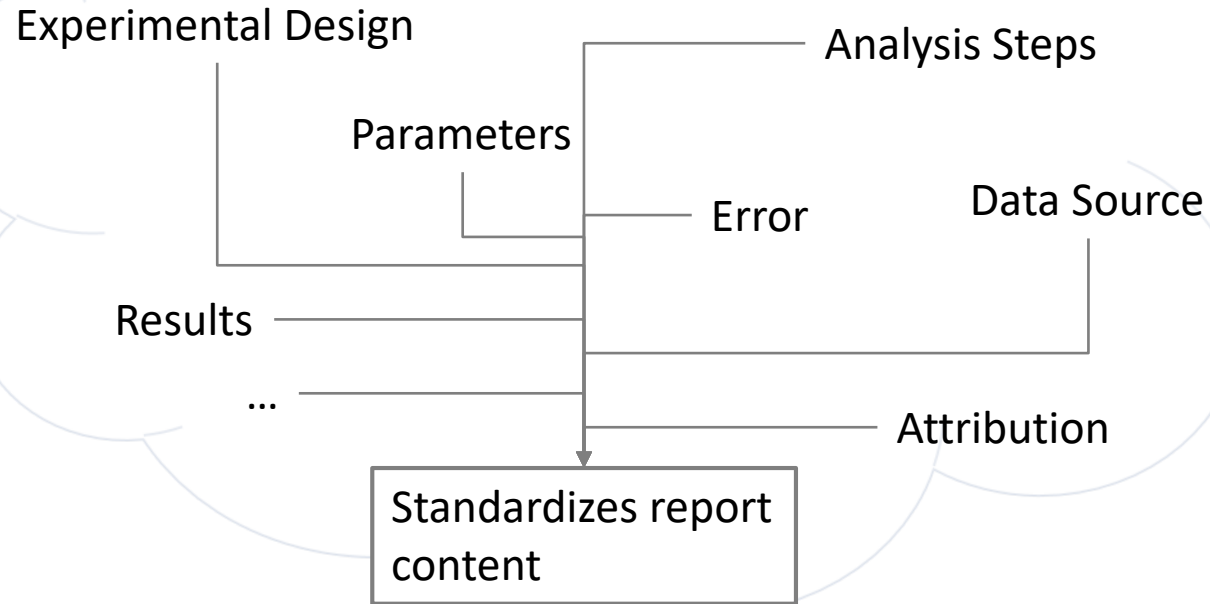
...

Attribution





# Solution: BioCompute



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** : fea7e938e6bdf9a2cfcb7fa02f5a5fc3973dccb0b03a64319e1ee29966a5b6b

### 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

IEEE 2791-2020 (“BioCompute”) 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](#)



This document has a comment period that ends in 24 days. (08/21/2020)

[SUBMIT A FORMAL COMMENT](#)

### 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



### 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:

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

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

Edit tags

Revision: 1 | Latest | Edit | Fork | Export | Run Workflow rev1

SPEC | WORKFLOW DIAGRAM

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



Projects | Data | Apps

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

Start Query From:

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

Workflow diagram showing: File (ADD FILTER) -> Data Format (Remove Filter) -> Experimental Strategy (Remove Filter) -> Disease Type (Remove Filter)



Main Home HIVE Portal Links

CensusScope

HMB21-2\_R1

Parameters | Progress | Results | Taxonomy Details | Taxonomy Help | Convergence | Phylogenetic Tree | Tree View | Table | Subtree | What's Next? | Alignment

Loading	Status
Building histogram	Done 100%
Preparing alignments	Done 100%
Visualizing alignments in track	Done 100%
Fetching alignments	Done 100%
Creating mutation heat diagram	Done 100%

Taxonomy Help | Taxonomy Details

Alcemy | BioProjectID | Name | Taxname | Parent | Rank | Taxonomy ID



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)





### Object Options

- eMail Object
- Derivation Chain
- Download Object

### Display Options

- Meta
- Provenance Domain
- Description Domain
- Execution Domain
- Io Domain
- Usability Domain
- Parametric Domain
- Error Domain
- Extension Domain

spec version <https://w3id.org/ieee/ieee-2791-schema/>

eTag ca34683b739b6c283adc89bd9bdcbaa5c5f1056037164a8b2934567955a60420

### Provenance Domain

Name	WGS Simulation of DUF1220 Regions			
Version	3.0			
License	<a href="https://opensource.org/licenses/MIT">https://opensource.org/licenses/MIT</a>			
Created	2020-08-30T11:00:52.937Z			
Modified	2020-08-30T11:00:52.937Z			
Name	Contribution	ORCID	Affiliation	eMail
David Astling	authoredBy	<a href="https://orcid.org/0000-0001-8179-0304">https://orcid.org/0000-0001-8179-0304</a>	University of Colorado	david.astling@example.com
Ilea Heft	authoredBy	<a href="https://orcid.org/0000-0002-7759-7007">https://orcid.org/0000-0002-7759-7007</a>	University of Colorado	ilea.heft@example.com
Kenneth Jones	authoredBy	None	University of Colorado	kenneth.jones@example.com
James Sikela	authoredBy	<a href="https://orcid.org/0000-0001-5820-2762">https://orcid.org/0000-0001-5820-2762</a>	University of Colorado	james.sikela@example.com
Jonathon Keeney	createdBy	None	GWU	keeneyjg@gwu.edu
Alex Nguyen	createdBy	None	UVA	tan5um@virginia.edu

### Usability Domain

Pipeline for identifying copy number of genetic sequences independent of the genes in which they occur, and with higher fidelity than existing methods. Approximately 25



### Object Options

- eMail Object
- Derivation Chain
- Download Object

### Display Options

- Meta
- Provenance Domain
- Description Domain
- Execution Domain
- Io Domain
- Usability Domain
- Parametric Domain
- Error Domain
- Extension Domain

**Etag:** ca34683b739b6c283adc89bd9bdcbaa5c5f1056037164a8b2934567955a60420

**Object\_id:** https://beta.portal.aws.biochemistry.gwu.edu/BCO\_7/1.0

**Spec\_version:** https://w3id.org/ieee/ieee-2791-schema/

**Usability\_domain+**

**Provenance\_domain+**

**Description\_domain-**

**Keywords+**

**Platform+**

**Pipeline\_steps-**

0+

1+

2-

**Name:** Spike-In Trim and Filter Reads

**Version:** 1.0.0

**Step\_number:** 3

**Input\_list+**

**Output\_list+**

**Prerequisite+**

**Description:** This script filters and trims reads down to 100 bp to remove low quality bases from the ends.



# 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