BioCompute Objects on the High-performance Integrated Virtual Environment

JONATHON KEENEY, PH.D

ASSISTANT RESEARCH PROFESSOR, GEORGE WASHINGTON UNIVERSITY

HIVE

BioCompute Objects

OncoMX

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

hive.aws.biochemistry.gwu.edu/dna.cgi?cmd=home

MAIN HOME DATASETS PORTAL LINKS HIVE INSTANCES METADATA CONTACT HELP LOGOUT JONATHON KEENEY NEWS 5

home help

all[98] folders[6] genomes[2] reads[5] annotations[1] files[37] computations[47] data-loading[32]

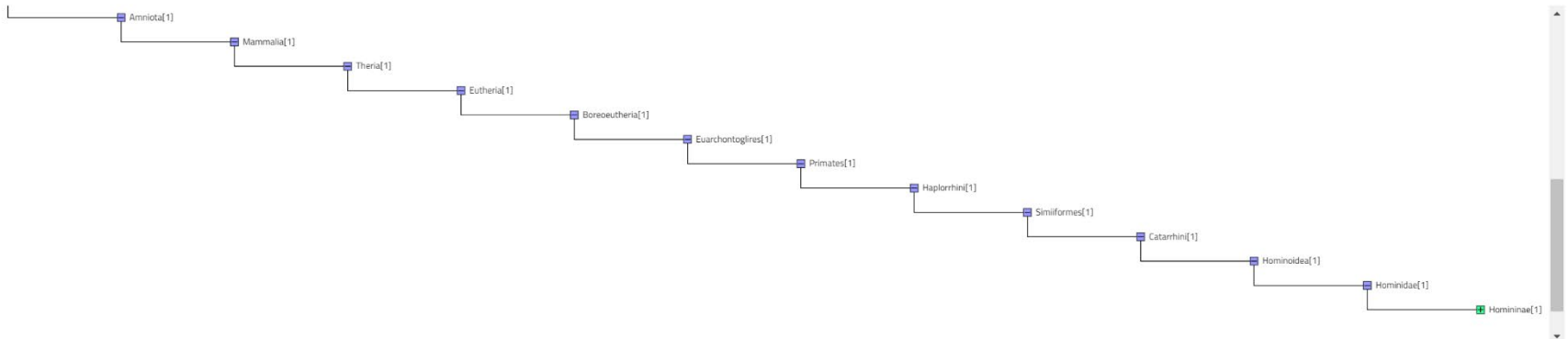
Cut Edit

- HIVE Space
- All objects
- Inbox
  - genome_assemblies_genome_fasta.tar
  - HCB
  - HumanGutDB_Epilepsy.zip
  - IgG-H1b.zip
  - MicrobiomeData
  - ProteinsBiomarkersClass
- Trash
  - sample

Cut Convert Export Details 1-50 of 98 50

ID	Summary	Created
5466	**Glymps for glyco-proteomics** 100%	4/6/2020
5355	IgG-H1b.zip **Folder**	4/6/2020
5353	glymps_protease_library_version-1.0.0.0.0.txt	4/6/2020
5352	glymps_mass_library_version-1.0.0.0.0.txt	4/6/2020
5351	glymps_mod_library_version-1.0.0.0.0.txt	4/6/2020
5350	IgG-H1b.zip **File Upload** 100%	4/6/2020
5349	glymps_mass_library_version-1.0.0.0.0.txt, glymps_mod_library_version-1.0.0.0.0.txt, glymps_protease_library_version-1.0.0.0.0.txt **File Upload** 100%	4/6/2020
5344	**GlycoPeptideSearch** 100%	4/5/2020
5343	iggs.fasta	4/4/2020
5342	iggs.fasta **File Upload** 100%	4/4/2020
4810	**MS-GF+** 100%	3/28/2020
4809	Example_Glycomics1 **MS-GF+** 100%	3/28/2020
4808	**MS-GF+** 100%	3/27/2020
4786	ProteinsBiomarkersClass **Folder**	3/27/2020
4785	'NM_004304.3.gb **ionAnnot type(s)**'	3/26/2020
4784	NM_004304.3.gb 1 **Genomic sequence(s)**	3/26/2020
4783	genbank: NM_004304.3 **Downloader Engine** 100%	3/26/2020
4734	MicrobiomeData **Folder**	3/22/2020
4733	11566.S22.fastq 166468 **Nucleotide read(s)**	3/22/2020
4732	Experimental **Downloader Engine** 81%	3/22/2020
4731	11566.S21.fastq 172731 **Nucleotide read(s)**	3/22/2020

ACGT sequences histogram positionalQC taxonomy sunburst codonQC details



HIVE Talks

Novel Approach for Identification of Defective Viral Genomes using NGS Data

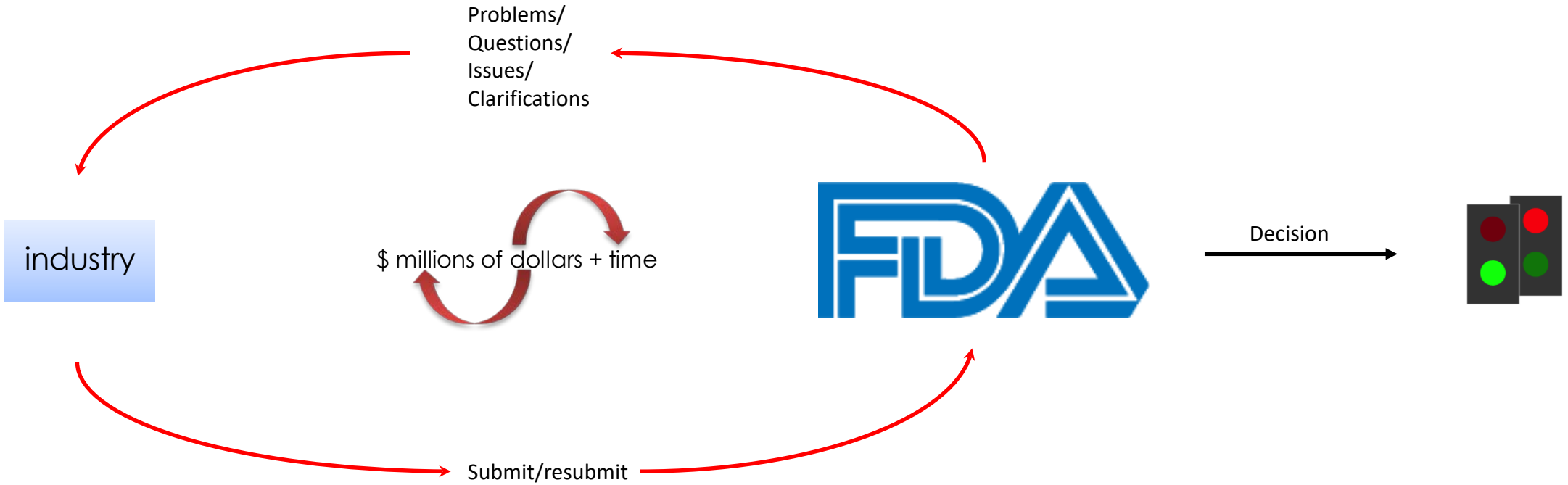
- Defective viral genomes (DVGs) are spontaneously generated by most viruses
- Some DVGs recognized as important triggers of antiviral innate immunity
- DVGs have implications for vaccine immunogenicity
- New HIVE tool: DVG-profiler
- Tool presentation, along with application in viral quasispecies analysis

Next talk: One-click RNAseq analysis

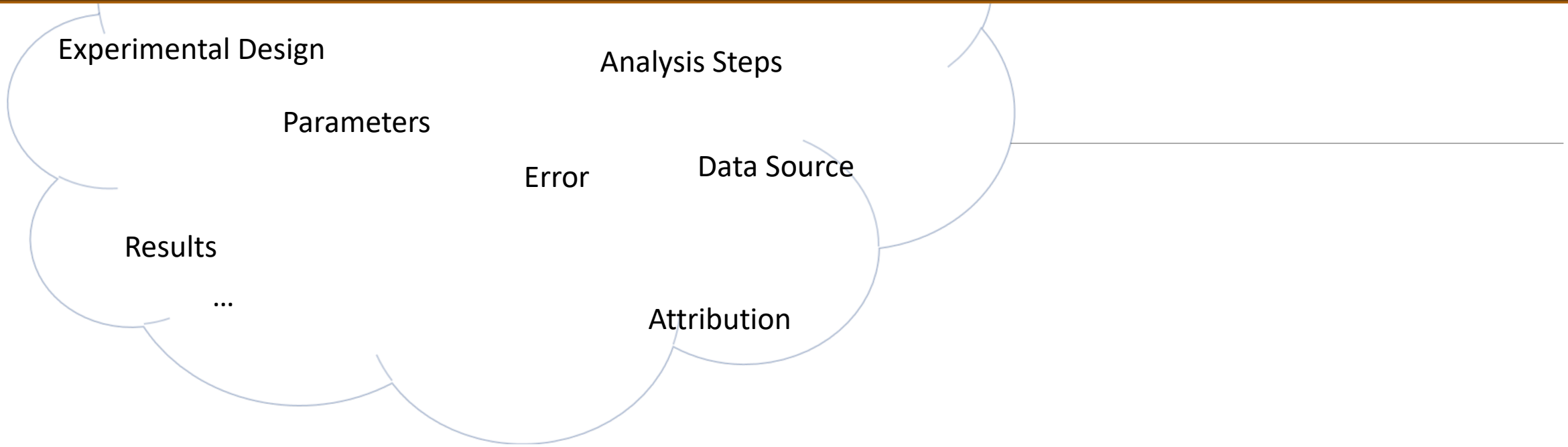
- Luis Santana-Quintero
- Date TBD

Konstantinos Karagiannis, Ph.D.

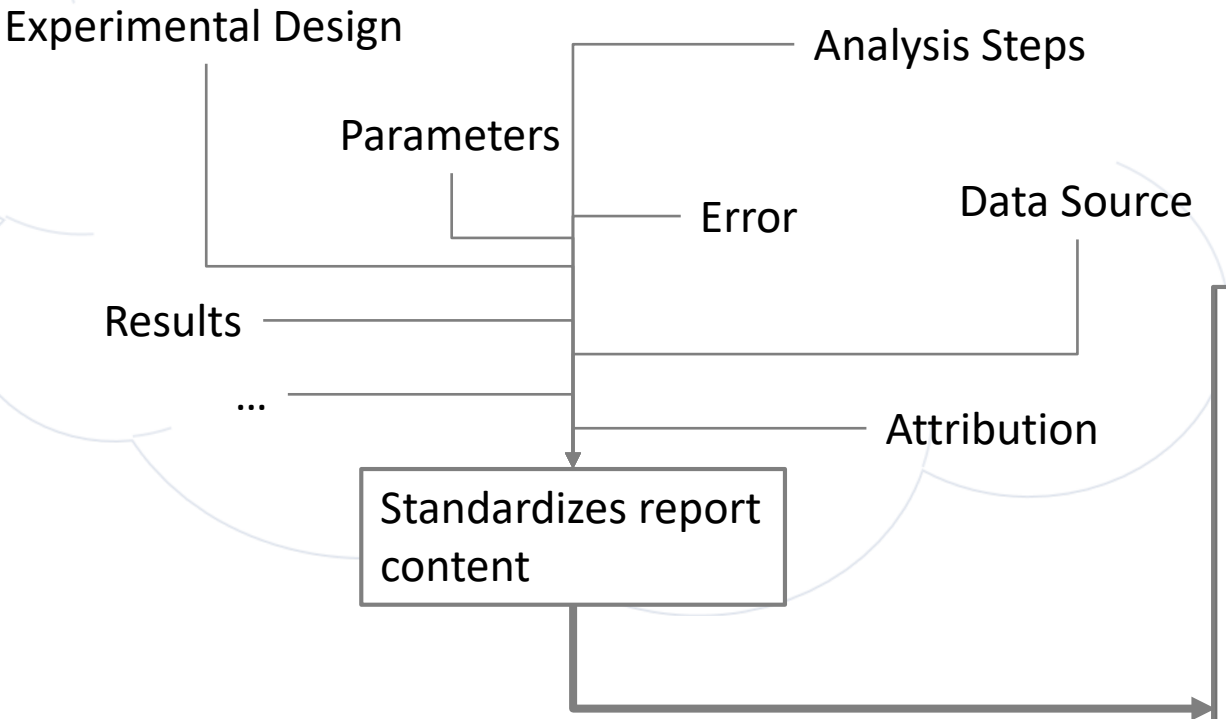
Wasted Time and Money



Solution: BioCompute



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]
```


Machine readability enables customized views

Metadata

object_id : https://beta.portal.aws.biochemistry.gwu.edu/bco/BCO_00016916
spec_version : https://w3id.org/ieee/ieee-2791-schema/
etag : fea7e938e6bdf9a2cfcba7fa02f5a5fc3973dccb0b03a64319e1ee29966a5b6b

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
    - 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) 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 Objects

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

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

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

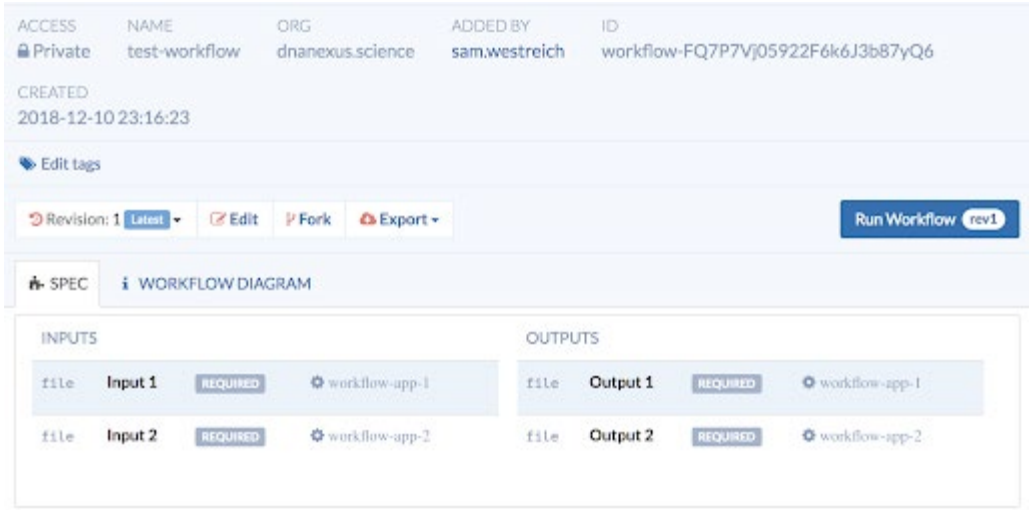
Printed version:
PDF

Publication Date:
07/22/2020

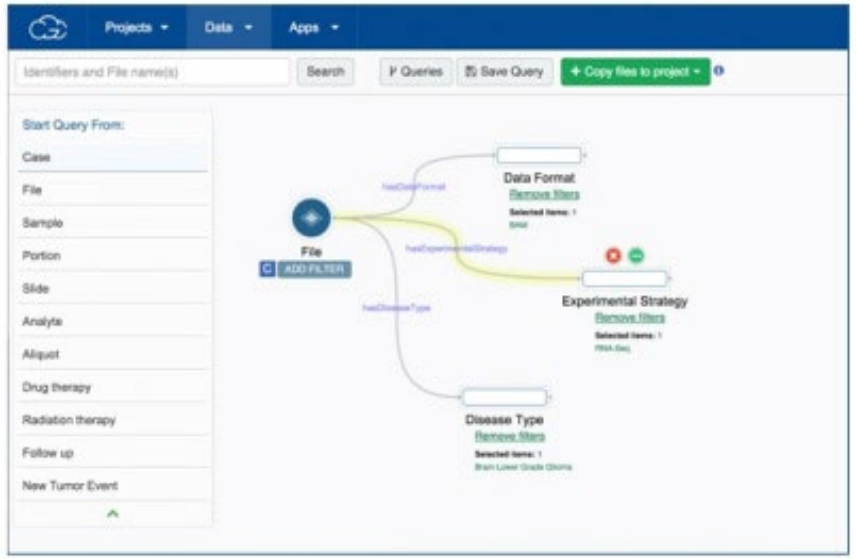
Agencies:
Food and Drug Administration

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

Comments Close:

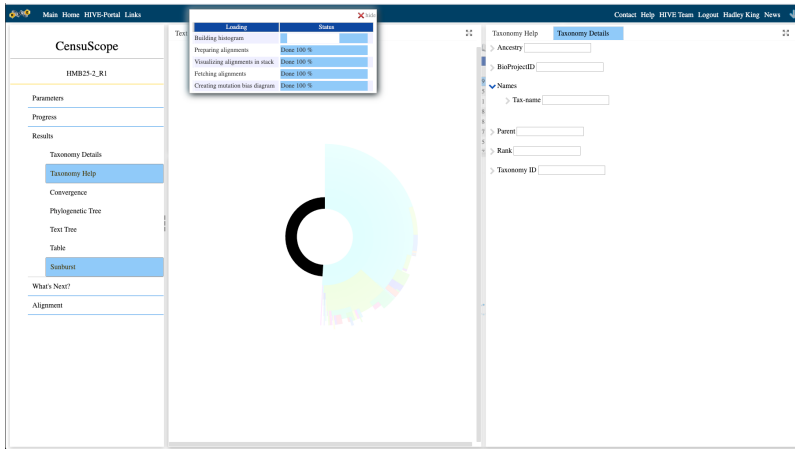


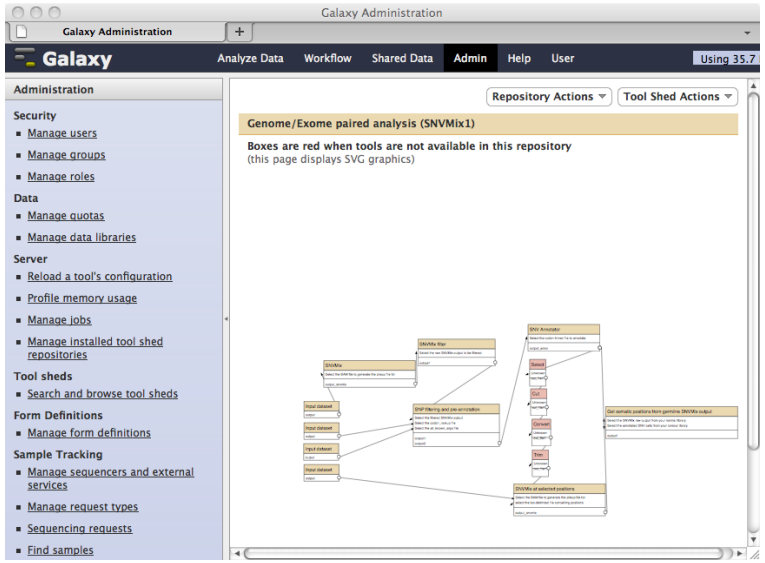
Seven Bridges



precisionFDA

DNAnexus
powering the genomics revolution™

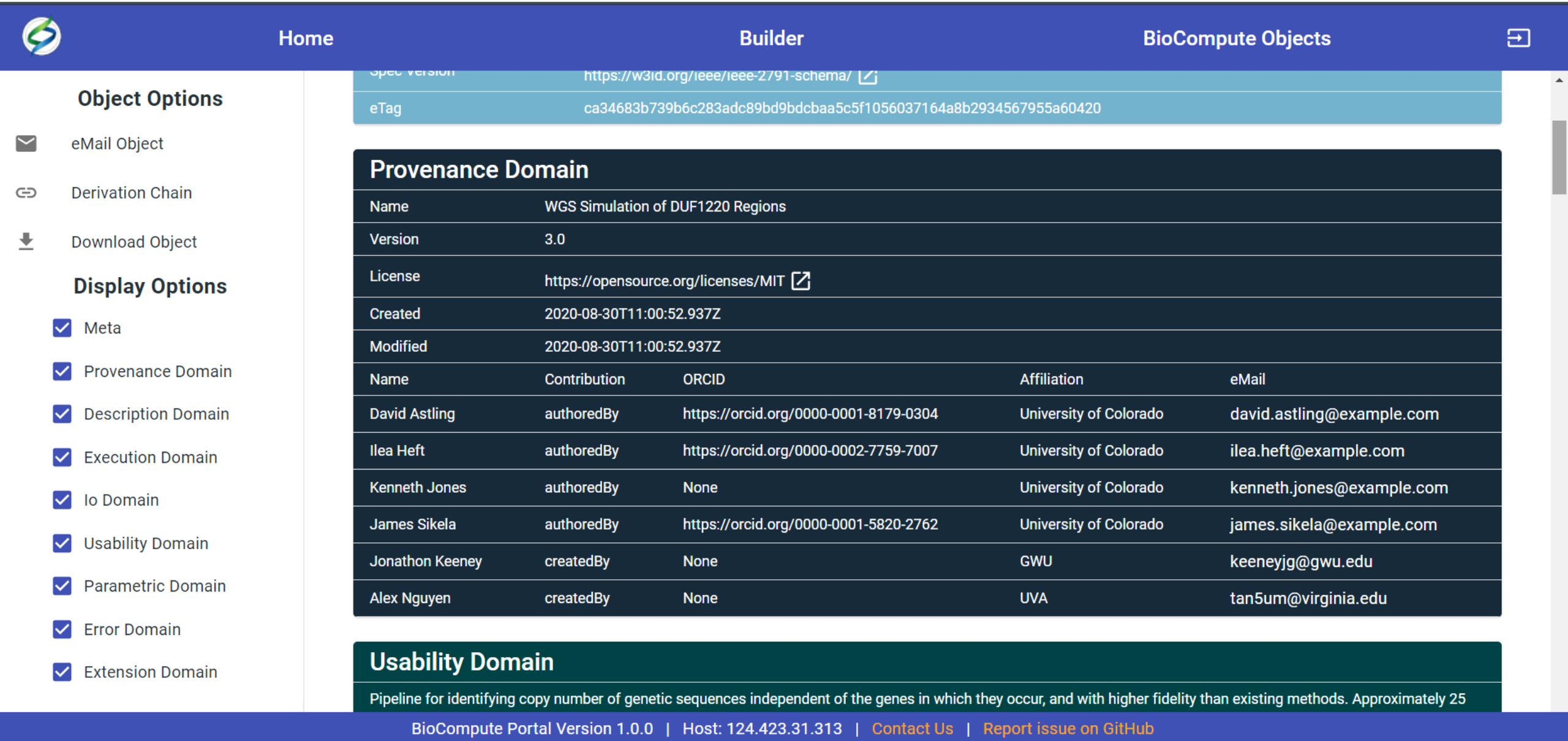




Galaxy PROJECT

HIVE

https://beta.portal.aws.biochemistry.gwu.edu/dashboard



https://beta.portal.aws.biochemistry.gwu.edu/dashboard

Acknowledgements and Contact

Raja Mazumder, Ph.D., PI

Professor

The George Washington University

mazumder@gwu.edu

Hadley King

Technical Lead

The George Washington University

Janisha Patel

Outreach Lead

The George Washington University

Jonathon Keeney, Ph.D., Co-I

Assistant Research Professor

The George Washington University

keeneyjg@gwu.edu

Chris Armstrong

Development Lead

The George Washington University

BioCompute Objects