BioCompute: A Standardized Method to Communicate Bioinformatic Workflow Information and Ease Organizational Burden

Jonathon Keeney, Ph.D.
Assistant Research Professor, George Washington University
Managing Director, BioCompute Executive Steering Committee

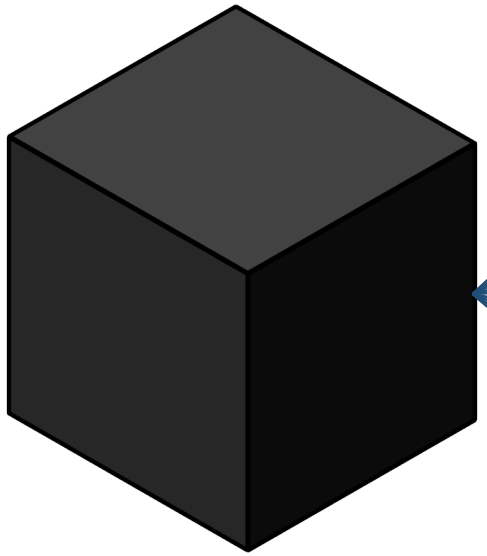
Webinar: Advanced Virus Detection Technologies Interest Group
April 2nd, 2020

keeneyjg@gwu.edu

NGS Data Flows

```
$ fastq-dump -X 2 SRR001666 --split-3
$ fastq-dump -X 2 SRR001666 --split-3
Read 2 spots for SRR001666
Written 2 spots for SRR001666
$ head SRR001666_1.fastq SRR001666_2.fastq
==> SRR001666_1.fastq <==
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
GTTCAGGGATACGACGTTTGTATTTTAAGAATCTGA
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBI

==> SRR001666_2.fastq <==
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
AAGTTACCCTTAACAACTTAAGGGTTTTCAAATAGA
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIDIIIIIII>IIIIII/
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
AGCAGAAGTCGATGATAATACGCGTCGTTTTATCAT
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
IIIIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```



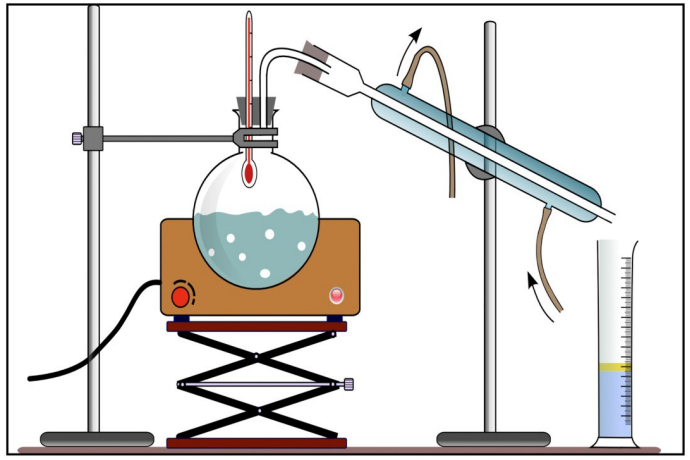
- Ancestry
- Cancer
- Microbiome
- Disease correlation
- Agriculture
- Synthetic biology
- Livestock
- Metagenomics
- Personalized medicine

BioCompute Objects

Challenge: Workflow Communication

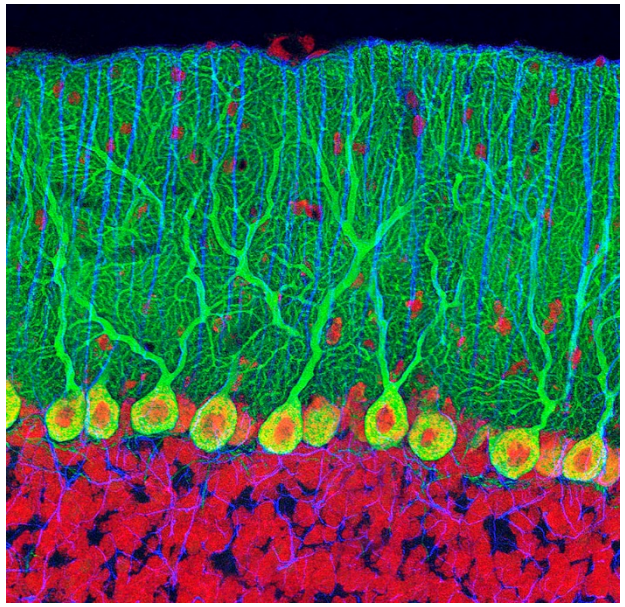
Input

Strain of animal
Tissue type
Protein Lot Number
Humidity
Microscope
…



conditions

Output



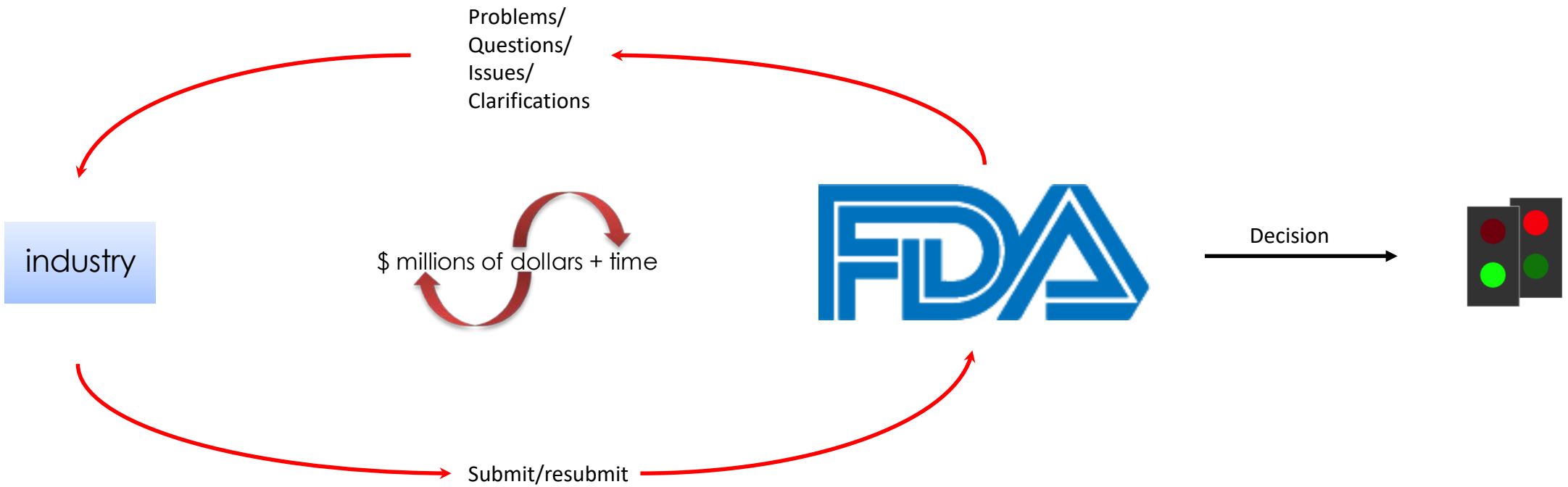
generalized experiment	inputs	parameters	outputs

```
> my_program -i input_file1 -parameter1 value1 -parameter2 value2 -o out_file
```

Analogy: wet lab experiments

BioCompute Objects

Wasted Time and Money



BioCompute Objects

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

A Solution Should…

- Be human readable: like a GenBank sequence record
- Be computer 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
- Have a way to be sure the information has not been altered: immutable

BioCompute Objects

Solution: BioCompute

- Standard for communicating genomic analysis workflows
- Acts like an envelope for entire pipeline
  - Can incorporate other standards (e.g. CWL, FHIR Genomics)
- Built with FDA
- Human and machine readable
  - Written in JSON
- Categorized by domains
- Adheres to F.A.I.R. principles
- Adaptable
- Preserves data provenance
- Unique IDs for versioning

BioCompute Objects

802.11 Analogy

BioCompute Objects

BioCompute Object

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/1xUY7WJNEZHyCgH5sYpxEuqAbtgVUUwgR2oc0IWhH28Y/edit#gid=1492026303

IO domain

Error Domain

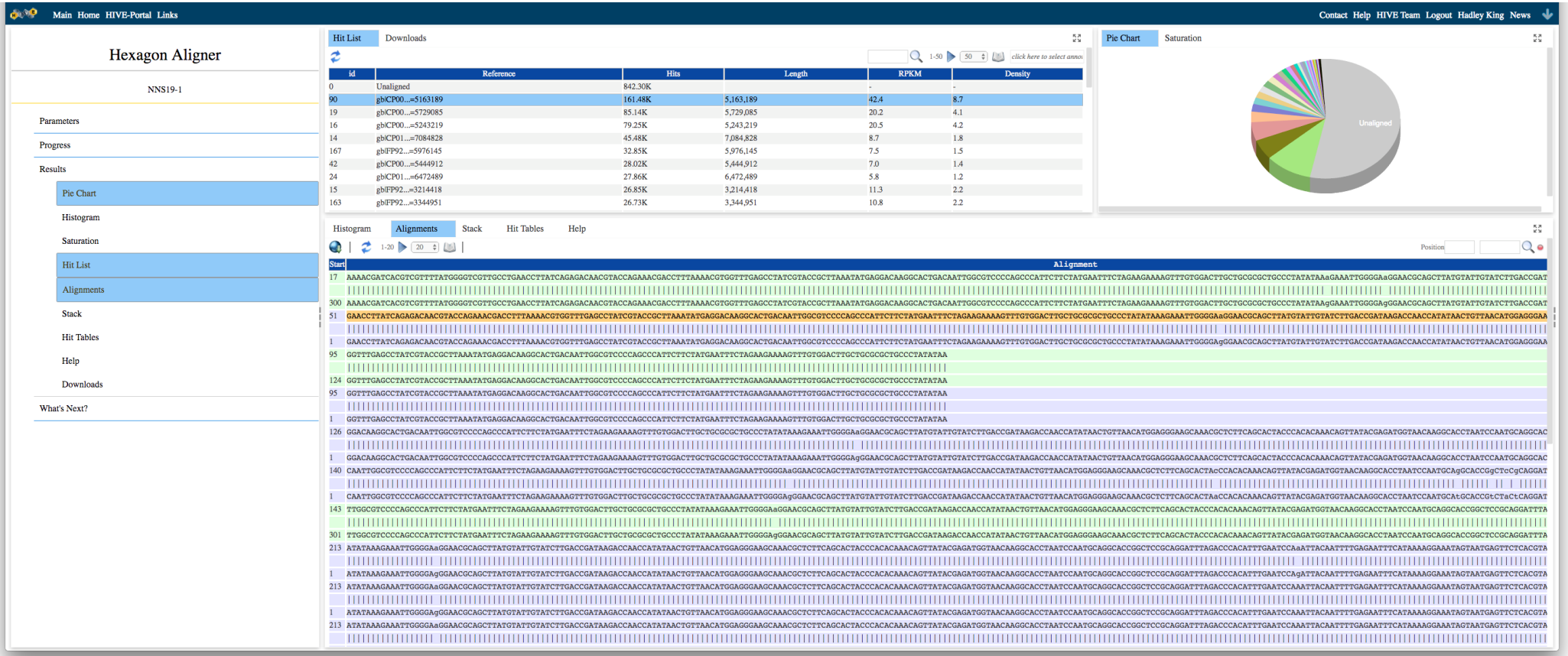
Error domain

BCO Example: HCV-1 drug resistance

- Goal: Identify SNPs, insertions, and deletions that correlate with reduced ledipasvir antiviral drug efficacy in Hepatitis C virus subtype 1
  - Genome sequencing data from a drug resistant cohort

BioCompute Objects

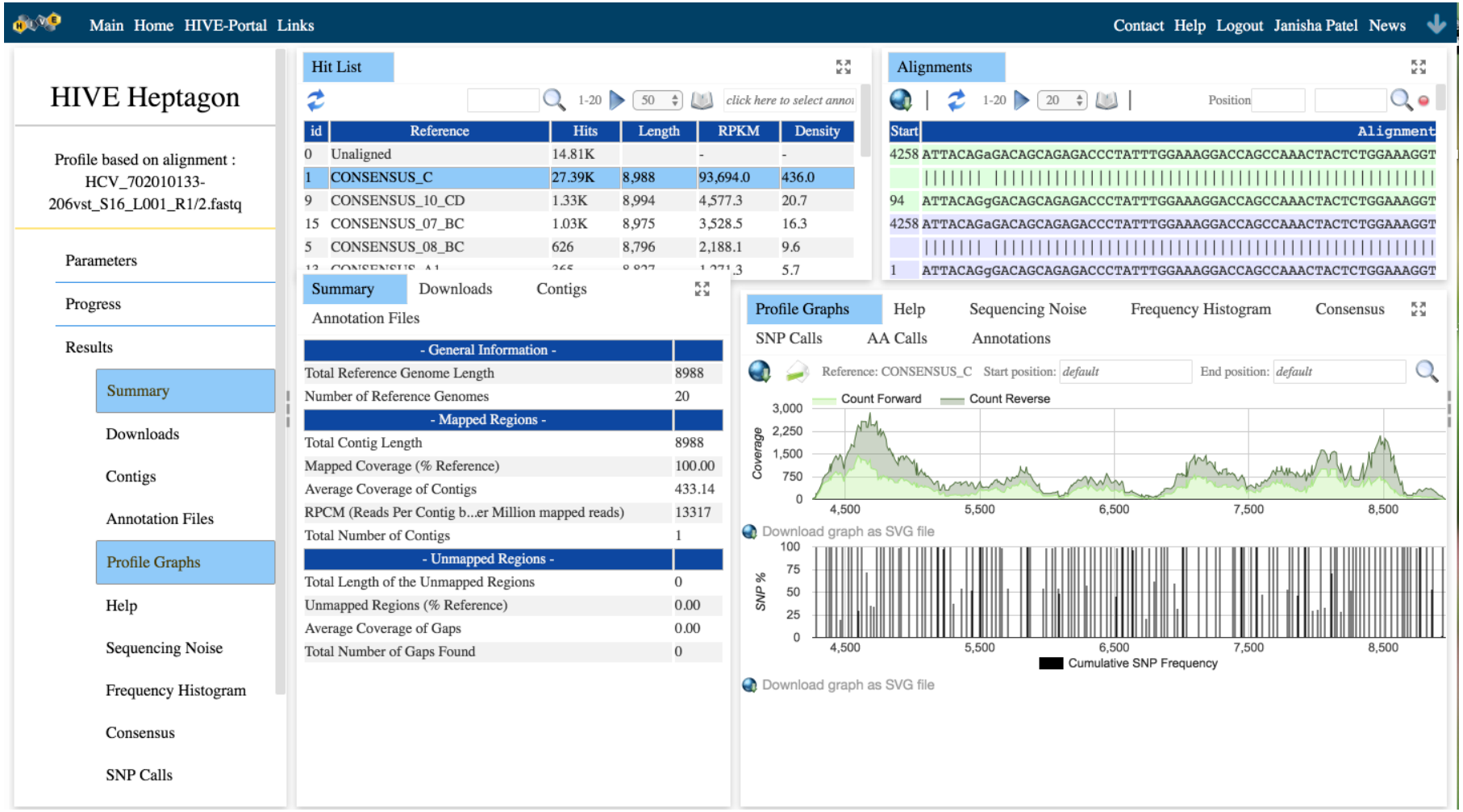
BCO Example: HCV-1 drug resistance



id	Reference	Hits	Length	RPKM	Density
0	Unaligned	842.30K		-	-
90	gb|CP00...=5163189	161.48K	5,163,189	42.4	8.7
19	gb|CP00...=5729085	85.14K	5,729,085	20.2	4.1
16	gb|CP00...=5243219	79.25K	5,243,219	20.5	4.2
14	gb|CP01...=7084828	45.48K	7,084,828	8.7	1.8
167	gb|FP92...=5976145	32.85K	5,976,145	7.5	1.5
42	gb|CP00...=5444912	28.02K	5,444,912	7.0	1.4
24	gb|CP01...=6472489	27.86K	6,472,489	5.8	1.2
15	gb|FP92...=3214418	26.85K	3,214,418	11.3	2.2
163	gb|FP92...=3344951	26.73K	3,344,951	10.8	2.2

BioCompute Objects

BCO Example: HCV-1 drug resistance



BioCompute Objects

```
{
   "object_id": 1270,
   "type": "antiviral_resistance_detection",
   "provenance_domain": {
       "name": "HCV1a ledipasvir resistance SNP detection",
       "structured_name": "HCV1a [taxID:31646] ledipasvir
[PubChem:67505836] resistance SNP [SO:0000694] detection",
       "version": "1.1",
       "digital_signature": "905d7fce3f3ac64c8ea86f058ca71658",
       "review": [
           {
               "status": "approved",
               "comment": "Approved by GW staff. Waiting for approval from
FDA Reviewer",
               "reviewer": {"type": "orcid", "orcid":
"http://orcid.org/0000-0003-1409-4549" }
           },
           {
               "status": "approved",
               "comment": "The revised BCO looks fine",
               "reviewer": {
                   "type": "entity",
                   "entity": {
                       "name": "Eric Donaldson",
                       "affiliation": "FDA",
                       "email": "Eric.Donaldson@fda.hhs.gov"
                   }
               }
           }
       ],
       "derived_from" : null,
       "obsolete" : null,
       "embargo" : {
           "start_time": null,
           "end_time": null
       },
       "created": "2017-01-24T09:40:17-0500",
       "modified": "2018-03-21T18:31:48-0400",
       "contributors": [
           {
               "type": "entity",
               "entity": {
                   "name": "Eric Donaldson",
                   "affiliation": "FDA",
                   "email": "Eric.Donaldson@fda.hhs.gov",
                   "contribution": "authoredBy"
               }
           },
           {
               "type": "orcid",
               "orcid": "http://orcid.org/0000-0003-1409-4549",
               "contribution": "createdBy"
           }
       ],
      "license":{
          "address": "https://creativecommons.org/licenses/by/4.0/",
          "access_time": "2017-01-24T09:40:17-0500"
      }
   },
   "usability_domain": [
       "Identify baseline single nucleotide polymorphisms SNPs [SO:0000694], insertions
[SO:0000667], and deletions [SO:0000045] that correlate with reduced ledipasvir
[PubChem:67505836] antiviral drug efficacy in Hepatitis C virus subtype 1 [taxID:31646]",
       "Identify treatment emergent amino acid substitutions [SO:0000048] that correlate
with antiviral drug treatment failure",
       "Determine whether the treatment emergent amino acid substitutions [SO:0000048]
identified correlate with treatment failure involving other drugs against the same
virus",
       "GitHub CWL example: https://github.com/mr-c/hive-cwl-
examples/blob/master/workflow/hive-viral-mutation-detection.cwl#L20"
   ],
   "extension_domain":{
       "github_extension": {
          "github_URL": "https://github.com/common-workflow-language/hive-cwl-examples",
          "gitHub_URI": {
            "pipeline_steps": {
               "tool": [
                   {
                       "step_number": "1",
                       "name": "HIVE-hexagon",
                       "description": "Alignment of reads to a set of references",
                       "version": "1.3",
                       "prerequisite": [
                           {
                               "name": "Hepatitis C virus genotype 1",
                               "URI": {
                                   "address":
"http://www.ncbi.nlm.nih.gov/nuccore/22129792",
                                   "access_time": "2017-01-24T09:40:17-0500"
                               }
                           },
                           {
                               "name": "Hepatitis C virus type 1b complete genome",
                               "URI": {
                                   "address":
"http://www.ncbi.nlm.nih.gov/nuccore/5420376",
                                   "access_time": "2017-01-24T09:40:17-0500"
                               }
                           },
                           {
                               "name": "Hepatitis C virus (isolate JFH-1) genomic RNA",
                               "URI": {
                                   "address":
"http://www.ncbi.nlm.nih.gov/nuccore/13122261",
                                   "access_time": "2017-01-24T09:40:17-0500"
                               }
                           },
                           {
                               "name": "Hepatitis C virus clone J8CF, complete genome",
                               "URI": {
                                   "address":
"http://www.ncbi.nlm.nih.gov/nuccore/386646758",
                                   "access_time": "2017-01-24T09:40:17-0500"
                               }
                           },
                           {
                               "name": "Hepatitis C virus S52 polyprotein gene",
                               "URI": {
                                   "address":
"http://www.ncbi.nlm.nih.gov/nuccore/295311559",
                                   "access_time": "2017-01-24T09:40:17-0500"
                               }
                           }
                       ],
                       "input_uri_list": [
                           {
                               "address":
"https://hive.biochemistry.gwu.edu/dna.cgi?cmd=objFile&ids=514683",
                               "access_time": "2017-01-24T09:40:17-0500"
                           },
                           {
                               "address":
"https://hive.biochemistry.gwu.edu/dna.cgi?cmd=objFile&ids=514682",
                               "access_time": "2017-01-24T09:40:17-0500"
                           }
                       ],
                       "output_uri_list": [
                           {
                               "address":
"https://hive.biochemistry.gwu.edudata/514769/allCount-aligned.csv",
                               "access_time": "2017-01-24T09:40:17-0500"
                           }
                       ]
                   },
                   {
                       "step_number": "2",
                       "name": "HIVE-heptagon",
                       "description": "variant calling",
                       "version": "1.3",
                       "prerequisites": null,
                       "input_uri_list": [
                           {
                               "address":
"https://hive.biochemistry.gwu.edudata/514769/dnaAccessionBased.csv",
                               "access_time": "2017-01-24T09:40:17-0500"
                           }
                       ],
                       "output_uri_list": [
                           {
                               "address":
"https://hive.biochemistry.gwu.edudata/514801/SNPProfile.csv",
                               "access_time": "2017-01-24T09:40:17-0500"
                           },
                           {
                               "address":
"https://hive.biochemistry.gwu.edudata/14769/allCount-aligned.csv",
                               "access_time": "2017-01-24T09:40:17-0500"
                           }
                       ]
                   }
               ]
           }
       },
       "execution_domain": {
           "script_type": "text",
           "script": "https://hive.biochemistry.gwu.edu/dna.cgi?cmd=home",
           "script_driver": "manual",
           "pipeline_version": "2.0",
           "platform": "hive",
           "software_prerequisites": [
               {
                   "name": "HIVE-hexagon",
                   "version": "babajanian.1",
                   "uri": {
                       "address":
"https://hive.biochemistry.gwu.edu/dna.cgi?cmd=dna-hexagon&cmdMode=-",
                       "access_time": "2017-01-24T09:40:17-0500"
                   }
               },
               {
                   "name": "HIVE-heptagon",
                   "version": "albinoni.2"
               }
           ],
           "domain_prerequisites": [
               {
                   "name": "access HIVE",
                   "url": "https://hive.biochemistry.gwu.edu/dna.cgi?cmd=login"
               },
               {
                   "name": "access to ftp",
                   "url": "ftp://:22/"
               },
               {
                   "name": "access to e-utils",
                   "url": "http://eutils.ncbi.nlm.nih.gov/entrez/eutils/"
               }
           ],
           "env_parameters": {
               "HOSTTYPE" :"x86_64-linux"
           }
       },
       "parametric_domain": {
           "HIVE-hexagon" : {
               "seed": "14",
               "minimum_match_len": "66",
               "divergence_threshold_percent": "30"
           },
           "HIVE-heptagon": {
               "minimum_coverage": "15",
               "freq_cutoff": "0.10"
           }
       },
       "io_domain": {
           "input_subdomain": {
               "subject": [
                   {
```

```
{
   "object_id": 1270,
   "type": "antiviral_resistance_detection",
   "provenance_domain": {
       "name": "HCV1a ledipasvir resistance SNP detection",
       "structured_name": "HCV1a [taxID:31646] ledipasvir
[PubChem:67505836] resistance SNP [SO:0000694] detection",
       "version": "1.1",
       "digital_signature": "905d7fce3f3ac64c8ea86f058ca71658",
       "review": [
           {
               "status": "approved",
               "comment": "Approved by GW staff. Waiting for approval from
FDA Reviewer",
"http://or
       ],
       "de
       "ob
       "em
       },
       "cr
       "mo
       "co
       ],
   "lice
       "
       "
       }
   },
   "usability_domain": [
       "Identify baseline single nucleotide polymorphisms SNPs [SO:0000694], insertions
[SO:0000667], and deletions [SO:0000045] that correlate with reduced ledipasvir
[PubChem:67505836] antiviral drug efficacy in Hepatitis C virus subtype 1 [taxID:31646]",
       "Identify treatment emergent amino acid substitutions [SO:0000048] that correlate
with antiviral drug treatment failure",
       "Determine whether the treatment emergent amino acid substitutions [SO:0000048]
identified correlate with treatment failure involving other drugs against the same
virus",
       "GitHub CWL example: https://github.com/mr-c/hive-cwl-
examples/blob/master/workflow/hive-viral-mutation-detection.cwl#L20"
   ],
   "extension_domain":{
       "github_extension": {
          "github_URL": "https://github.com/common-workflow-language/hive-cwl-examples",
          "gitHub_URI": {
```

```
        }
    },
    "usability_domain": [
        "Identify baseline single nucleotide polymorphisms SNPs [SO:0000694], insertions
[SO:0000667], and deletions [SO:0000045] that correlate with reduced ledipasvir
[PubChem:67505836] antiviral drug efficacy in Hepatitis C virus subtype 1 [taxID:31646]",
        "Identify treatment emergent amino acid substitutions [SO:0000048] that correlate
with antiviral drug treatment failure",
        "Determine whether the treatment emergent amino acid substitutions [SO:0000048]
identified correlate with treatment failure involving other drugs against the same
virus",
        "GitHub CWL example: https://github.com/mr-c/hive-cwl-
examples/blob/master/workflow/hive-viral-mutation-detection.cwl#L20"
    ],
    "extension_domain":{
        "github_extension": {
            "github_URL": "https://github.com/common-workflow-language/hive-cwl-examples",
            "gitHub_URI": {
```

```
"pipeline_steps": {
    "tool": [
        {
            "step_number": "1",
            "name": "HIVE-hexagon",
            "description": "Alignment of reads to a set of references",
            "version": "1.3",
            "prerequisite": [
                {
                    "name": "Hepatitis C virus genotype 1",
                    "URI": {
                        "address":
"http://www.ncbi.nlm.nih.gov/nuccore/22129792",
                        "access_time": "2017-01-24T09:40:17-0500"
                    }
                }
            ],
            "output_uri_list": [
                {
                    "address":
"https://hive.biochemistry.gwu.edudata/514769/allCount-aligned.csv",
                    "access_time": "2017-01-24T09:40:17-0500"
                }
            ]
        },
        {
            "step_number": "2",
            "name": "HIVE-heptagon",
            "description": "variant calling",
            "version": "1.3",
            "prerequisites": null,
            "input_uri_list": [
```

```
{
                    "address":
"https://hive.biochemistry.gwu.edudata/514769/dnaAccessionBased.csv",
                    "access_time": "2017-01-24T09:40:17-0500"
                }
            ],
            "output_uri_list": [
                {
                    "address":
"https://hive.biochemistry.gwu.edudata/514801/SNPProfile.csv",
                    "access_time": "2017-01-24T09:40:17-0500"
                },
                {
                    "address":
ry.gwu.edudata/14769/allCount-aligned.csv",
                    "access_time": "2017-01-24T09:40:17-0500"
ext",
//hive.biochemistry.gwu.edu/dna.cgi?cmd=home",
"manual",
": "2.0",
",
isites": [
HIVE-hexagon",
: "babajanian.1",
ess":
ry.gwu.edu/dna.cgi?cmd=dna-hexagon&cmdMode=-",
ss_time": "2017-01-24T09:40:17-0500"
HIVE-heptagon",
: "albinoni.2"
ites": [
access HIVE",
ttps://hive.biochemistry.gwu.edu/dna.cgi?cmd=login"
access to ftp",
tp://:22/"
access to e-utils",
ttp://eutils.ncbi.nlm.nih.gov/entrez/eutils/"
                {
x86_64-linux"
            }
        },
        "parametric_domain": {
            "HIVE-hexagon" : {
                "seed": "14",
                "minimum_match_len": "66",
                "divergence_threshold_percent": "30"
            },
            "HIVE-heptagon": {
                "minimum_coverage": "15",
                "freq_cutoff": "0.10"
            }
        },
        "io_domain": {
            "input_subdomain": {
                "subject": [
                    {
```

```
{
   "object_id": 1270,
   "type": "antiviral_resistance_detection",
   "provenance_domain": {
       "name": "HCV1a ledipasvir resistance SNP detection",
       "structured_name": "HCV1a [taxID:31646] ledipasvir
[PubChem:67505836] resistance SNP [SO:0000694] detection",
       "version": "1.1",
       "digital_signature": "905d7fce3f3ac64c8ea86f058ca71658",
       "review": [
           {
               "status": "approved",
               "comment": "Approved by GW staff. Waiting for approval from
FDA Reviewer",
               "reviewer": {"type": "orcid", "orcid":
"http://orcid.org/0000-0003-1409-4549" }
           },
           {
               "status": "approved",
               "comment": "The revised BCO looks fine",
               "reviewer": {
                   "type": "entity",
                   "entity": {
                       "name": "Eric Donaldson",
                       "affiliation": "FDA",
                       "email": "Eric.Donaldson@fda.hhs.gov"
                   }
               }
           }
       ],
       "derived_from" : null,
       "obsolete" : null,
       "embargo" : {
           "start_time": null,
           "end_time": null
       },
       "created": "2017-01-24T09:40:17-0500",
       "modified": "2018-03-21T18:31:48-0400",
       "contributors": [
           {
               "type": "entity",
               "entity": {
                   "name": "Eric Donaldson",
                   "affiliation": "FDA",
                   "email": "Eric.Donaldson@fda.hhs.gov",
                   "contribution": "authoredBy"
               }
           },
           {
               "type": "orcid",
               "orcid": "http://orcid.org/0000-0003-1409-4549",
               "contribution": "createdBy"
           }
       ],
      "license":{
          "address": "https://creativecommons.org/licenses/by/4.0/",
          "access_time": "2017-01-24T09:40:17-0500"
      }
   },
   "usability_domain": [
       "Identify baseline single nucleotide polymorphisms SNPs [SO:0000694], insertions
[SO:0000667], and deletions [SO:0000045] that correlate with reduced ledipasvir
[PubChem:67505836] antiviral drug efficacy in Hepatitis C virus subtype 1 [taxID:31646]",
       "Identify treatment emergent amino acid substitutions [SO:0000048] that correlate
with antiviral drug treatment failure",
       "Determine whether the treatment emergent amino acid substitutions [SO:0000048]
identified correlate with treatment failure involving other drugs against the same
virus",
       "GitHub CWL example: https://github.com/mr-c/hive-cwl-
examples/blob/master/workflow/hive-viral-mutation-detection.cwl#L20"
   ],
   "extension_domain":{
       "github_extension": {
          "github_URL": "https://github.com/common-workflow-language/hive-cwl-examples",
          "gitHub_URI": {
```

```
       "pipeline_steps": {
          "tool": [
              {
                  "step_number": "1",
                  "name": "HIVE-hexagon",
                  "description": "Alignment of reads to a set of references",
                  "version": "1.3",
                  "prerequisite": [
                      {
                          "name": "Hepatitis C virus genotype 1",
                          "URI": {
                              "address":
'http://www.ncbi.nlm.nih.gov/nuccore/22129792",
                              "access_time": "2017-01-24T09:40:17-0500"
                          }
```

```
 {
                              "address":
"https://hive.biochemistry.gwu.edudata/514769/dnaAccessionBased.csv",
                              "access_time": "2017-01-24T09:40:17-0500"
                          }
                      ],
                      "output_uri_list": [
                          {
                              "address":
"https://hive.biochemistry.gwu.edudata/514801/SNPProfile.csv",
                              "access_time": "2017-01-24T09:40:17-0500"
                          },
                          {
                              "address":
"https://hive.biochemistry.gwu.edudata/14769/allCount-aligned.csv",
```

```
        "pipeline_steps": {
           "tool": [
               {
                   "step_number": "1",
                   "name": "HIVE-hexagon",
                   "description": "Alignment of reads to a set of references",
                   "version": "1.3",
                   "prerequisite": [
                       {
                           "name": "Hepatitis C virus genotype 1",
                           "URI": {
                               "address":
"http://www.ncbi.nlm.nih.gov/nuccore/22129792",
                               "access_time": "2017-01-24T09:40:17-0500"
                           }
                       },
                       {
                           "name": "Hepatitis C virus type 1b complete genome",
                           "URI": {
                               "address":
"http://www.ncbi.nlm.nih.gov/nuccore/5420376",
                               "access_time": "2017-01-24T09:40:17-0500"
                           }
                       },
                       {
                           "name": "Hepatitis C virus (isolate JFH-1) genomic
```

```
{
   "object_id": 1270,
   "type": "antiviral_resistance_detection",
   "provenance_domain": {
       "name": "HCV1a ledipasvir resistance SNP detection",
       "structured_name": "HCV1a [taxID:31646] ledipasvir
[PubChem:67505836] resistance SNP [SO:0000694] detection",
       "version": "1.1",
       "digital_signature": "905d7fce3f3ac64c8ea86f058ca71658",
       "review": [
           {
               "status": "approved",
               "comment": "Approved by GW staff. Waiting for approval from
FDA Reviewer",
               "reviewer": {"type": "orcid", "orcid":
"http://orcid.org/0000-0003-1409-4549" }
           },
           {
               "status": "approved",
               "comment": "The revised BCO looks fine",
               "reviewer": {
                   "type": "entity",
                   "entity": {
                       "name": "Eric Donaldson",
                       "affiliation": "FDA",
                       "email": "Eric.Donaldson@fda.hhs.gov"
                   }
               }
           }
       ],
       "derived_f
       "obsolete"
       "embargo"
           "start
           "end_t
       },
       "created":
       "modified"
       "contribut
           {
               "t
               "e
               }
           },
           {
               "t
               "c
               "c
           }
       ],
      "license":{
          "address"
          "access_t
       }
   },
   "usability_domain
       "Identify bas
[SO:0000667], and de
[PubChem:67505836] a
       "Identify tre
with antiviral drug
       "Determine wh
identified correlate
virus",
       "GitHub CWL e
examples/blob/master
   ],
   "extension_domain
       "github_exten
          "github_U
          "gitHub_U
"pipeline_steps": {
   "tool": [
       {
           "step_number": "1",
           "name": "HIVE-hexagon",
           "description": "Alignment of reads to a set of references",
           "version": "1.3",
           "prerequisite": [
               {
                   "name": "Hepatitis C virus genotype 1",
                   "URI": {
                       "address":
"http://www.ncbi.nlm.nih.gov/nuccore/22129792",
                       "access_time": "2017-01-24T09:40:17-0500"
                   }
               },
               {
                   "name": "Hepatitis C virus type 1b complete genome",
                   "URI": {
                       "address":
"http://www.ncbi.nlm.nih.gov/nuccore/5420376",
                       "access_time": "2017-01-24T09:40:17-0500"
                   }
               },
               {
                   "name": "Hepatitis C virus (isolate JFH-1) genomic RNA",
                   "URI": {
    },
    "execution_domain": {
        "script_type": "text",
        "script": "https://hive.biochemistry.gwu.edu/dna.cgi?cmd=home",
        "script_driver": "manual",
        "pipeline_version": "2.0",
        "platform": "hive",
        "software_prerequisites": [
            {
                "name": "HIVE-hexagon",
                "version": "babajanian.1",
                "uri": {
                    "address":
"https://hive.biochemistry.gwu.edu/dna.cgi?cmd=dna-hexagon&cmdMode=-",
                    "access_time": "2017-01-24T09:40:17-0500"
                }
            },
            {
                "name": "HIVE-heptagon",
"                "version": "albinoni.2"
```

```
                        "address":
"https://hive.biochemistry.gwu.edudata/514769/dnaAccessionBased.csv",
                        "access_time": "2017-01-24T09:40:17-0500"
                    }
                ],
                "output_uri_list": [
                    {
                        "address":
"https://hive.biochemistry.gwu.edudata/514801/SNPProfile.csv",
                        "access_time": "2017-01-24T09:40:17-0500"
                    },
                    {
                        "address":
"https://hive.biochemistry.gwu.edudata/14769/allCount-aligned.csv",
                        "access_time": "2017-01-24T09:40:17-0500"
                    }
                ]
            }
        ]
    },
    "execution_domain": {
        "script_type": "text",
        "script": "https://hive.biochemistry.gwu.edu/dna.cgi?cmd=home",
        "script_driver": "manual",
        "pipeline_version": "2.0",
        "platform": "hive",
        "software_prerequisites": [
            {
                "name": "HIVE-hexagon",
                "version": "babajanian.1",
                "uri": {
                    "address":
s://hive.biochemistry.gwu.edu/dna.cgi?cmd=dna-hexagon&cmdMode=-",
                    "access_time": "2017-01-24T09:40:17-0500"
                }
            },
            {
                "name": "HIVE-heptagon",
                 "version": "albinoni.2"
            }
        ],
        "domain_prerequisites": [
            {
                "name": "access HIVE",
                "url": "https://hive.biochemistry.gwu.edu/dna.cgi?cmd=login"
            },
            {
                "name": "access to ftp",
                "url": "ftp://:22/"
            },
            {
                "name": "access to e-utils",
                "url": "http://eutils.ncbi.nlm.nih.gov/entrez/eutils/"
            }
        ],
        "env_parameters": {
            "HOSTTYPE" :"x86_64-linux"
        }

    arametric_domain": {
        "HIVE-hexagon" : {
            "seed": "14",
            "minimum_match_len": "66",
            "divergence_threshold_percent": "30"
        },
        "HIVE-heptagon": {
            "minimum_coverage": "15",
            "freq_cutoff": "0.10"
        }

    o_domain": {
        "input_subdomain": {
            "subject": [
                {
```

```
{
   "object_id": 1270,
   "type": "antiviral_resistance_detection",
   "provenance_domain": {
       "name": "HCV1a ledipasvir resistance SNP detection",
       "structured_name": "HCV1a [taxID:31646] ledipasvir
[PubChem:67505836] resistance SNP [SO:0000694] detection",
       "version": "1.1",
       "digital_signature": "905d7fce3f3ac64c8ea86f058ca71658",
       "review": [
           {
               "status": "approved",
               "comment": "Approved by GW staff. Waiting for approval from
FDA Reviewer",
               "reviewer": {"type": "orcid", "orcid":
"http://orcid.org/0000-0003-1409-4549" }
           },
           {
               "status": "approved",
               "comment": "The revised BCO looks fine",
               "reviewer": {
                   "type": "entity",
                   "entity": {
                       "name": "Eric Donaldson",
                       "affiliation": "FDA",
                       "email": "Eric.Donaldson@fda.hhs.gov"
                   }
               }
           }
       ],
       "derived_from" : null,
       "obsolete" : null,
       "embargo" : {
           "start_time": null,
           "end
       },
       "created
       "modifie
       "contrib
           {
           },
           {
           }
       ],
      "license":{
          "addres
          "access
       }
   },
   "usability_doma
       "Identify b
[SO:0000667], and deletions [SO:0000045] that correlate with reduced ledipasvir
[PubChem:67505836] antiviral drug efficacy in Hepatitis C virus subtype 1 [taxID:31646]",
       "Identify treatment emergent amino acid substitutions [SO:0000048] that correlate
with antiviral drug treatment failure",
       "Determine whether the treatment emergent amino acid substitutions [SO:0000048]
identified correlate with treatment failure involving other drugs against the same
virus",
       "GitHub CWL example: https://github.com/mr-c/hive-cwl-
examples/blob/master/workflow/hive-viral-mutation-detection.cwl#L20"
   ],
   "extension_domain":{
       "github_extension": {
          "github_URL": "https://github.com/common-workflow-language/hive-cwl-examples",
          "gitHub_URI": {
```

```
"pipeline_steps": {
   "tool": [
       {
           "step_number": "1",
           "name": "HIVE-hexagon",
           "description": "Alignment of reads to a set of references",
           "version": "1.3",
           "prerequisite": [
               {
                   "name": "Hepatitis C virus genotype 1",
                   "URI": {
                       "address":
"http://www.ncbi.nlm.nih.gov/nuccore/22129792",
                       "access_time": "2017-01-24T09:40:17-0500"
                   }
               },
               {
                   "name": "Hepatitis C virus type 1b complete genome",
                   "URI": {
                       "address":
"http://www.ncbi.nlm.nih.gov/nuccore/5420376",
                       "access_time": "2017-01-24T09:40:17-0500"
                   }
               },
               {
                   "name": "Hepatitis C virus (isolate JFH-1) genomic RNA",
                   "URI": {
                       "address":
"http://www.ncbi.nlm.nih.gov/nuccore/13122261",
                       "access_time": "2017-01-24T09:40:17-0500"
                   }
               },
               {
                   "name": "Hepatitis C virus clone J8CF, complete genome",
```

```
    "parametric_domain": {
        "HIVE-hexagon" : {
            "seed": "14",
            "minimum_match_len": "66",
            "divergence_threshold_percent": "30"
        },
        "HIVE-heptagon": {
            "minimum_coverage": "15",
            "freq_cutoff": "0.10"
        }
    },
```

```
                    "output_uri_list" : [
                        {
                            "address":
"https://hive.biochemistry.gwu.edudata/514769/allCount-aligned.csv",
                            "access_time": "2017-01-24T09:40:17-0500"
                        }
                    ]
                },
                {
                    "step_number": "2",
                    "name": "HIVE-heptagon",
                    "description": "variant calling",
                    "version": "1.3",
                    "prerequisites": null,
                    "input_uri_list": [
```

```
                        {
                            "address":
"https://hive.biochemistry.gwu.edudata/514769/dnaAccessionBased.csv",
                            "access_time": "2017-01-24T09:40:17-0500"
                        }
                    ],
                    "output_uri_list": [
                        {
                            "address":
"https://hive.biochemistry.gwu.edudata/514801/SNPProfile.csv",
                            "access_time": "2017-01-24T09:40:17-0500"
                        },
                        {
                            "address":
"https://hive.biochemistry.gwu.edudata/14769/allCount-aligned.csv",
                            "access_time": "2017-01-24T09:40:17-0500"
                        }
                    ]
                }
            ]
        }
    },
    "execution_domain": {
        "script_type": "text",
        "script": "https://hive.biochemistry.gwu.edu/dna.cgi?cmd=home",
        "script_driver": "manual",
        "pipeline_version": "2.0",
        "platform": "hive",
        "software_prerequisites": [
            {
                "name": "HIVE-hexagon",
                "version": "babajanian.1",
                "uri": {
                    "address":
ttps://hive.biochemistry.gwu.edu/dna.cgi?cmd=dna-hexagon&cmdMode=-",
                    "access_time": "2017-01-24T09:40:17-0500"
                }
            },
            {
                "name": "HIVE-heptagon",
                 "version": "albinoni.2"
            }
        ],
        "domain_prerequisites": [
            {
                "name": "access HIVE",
                "url": "https://hive.biochemistry.gwu.edu/dna.cgi?cmd=login"
            },
            {
                "name": "access to ftp",
                "url": "ftp://:22/"
            },
            {
                "name": "access to e-utils",
                "url": "http://eutils.ncbi.nlm.nih.gov/entrez/eutils/"
            }
        ],
        "env_parameters": {
            "HOSTTYPE" :"x86_64-linux"
        }
    },
    "parametric_domain": {
        "HIVE-hexagon" : {
            "seed": "14",
            "minimum_match_len": "66",
            "divergence_threshold_percent": "30"
        },
        "HIVE-heptagon": {
            "minimum_coverage": "15",
            "freq_cutoff": "0.10"
        }
    },
    "io_domain": {
        "input_subdomain": {
            "subject": [
                {
```

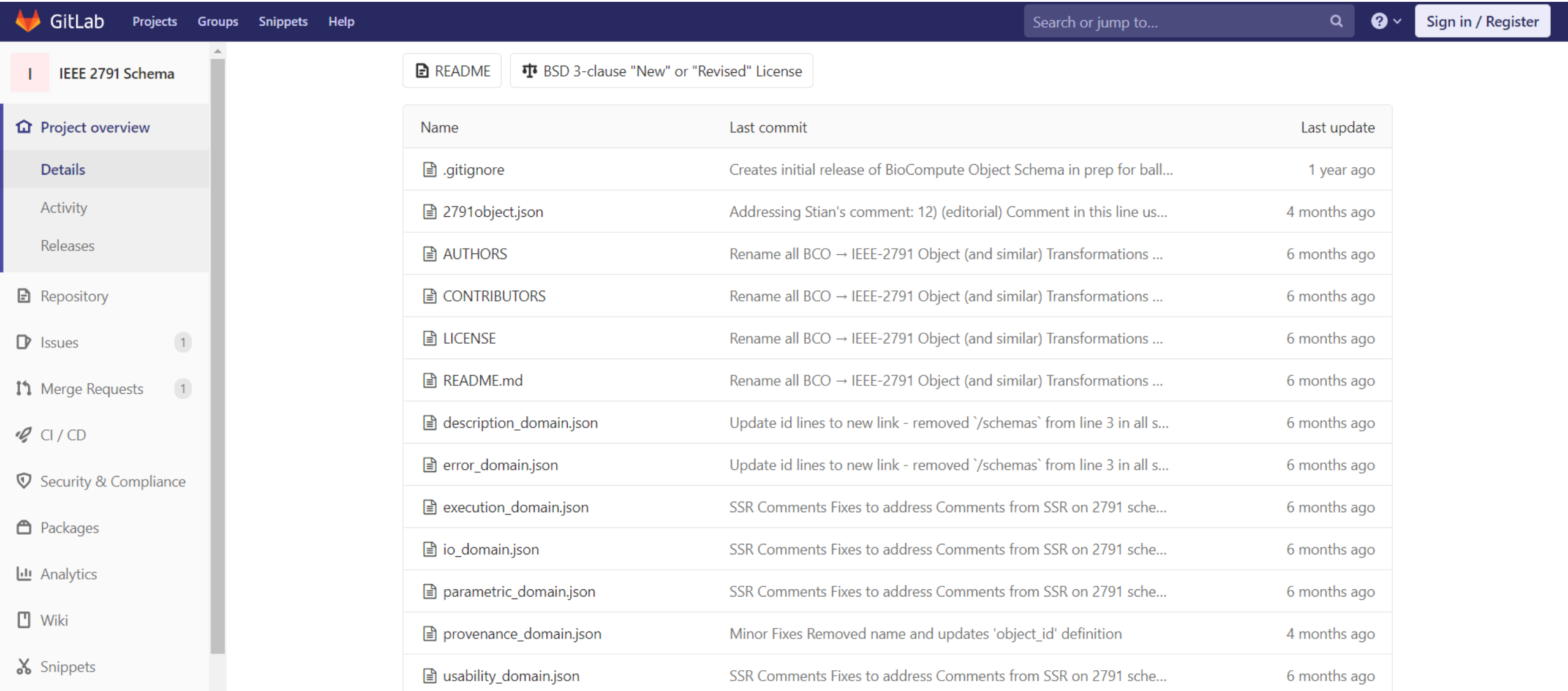
Key Aspects of a BCO

- Abstract away workflow based on commonalities
  - Platform/tool/protocol independent
- Usability Domain
  - Free text description
- Data provenance
  - Data manifest, track files from beginning to end
- Validation Kit
  - Error Domain + IO Domain
- Extensible
  - Extension Domain
  - Open source repository

BioCompute Objects

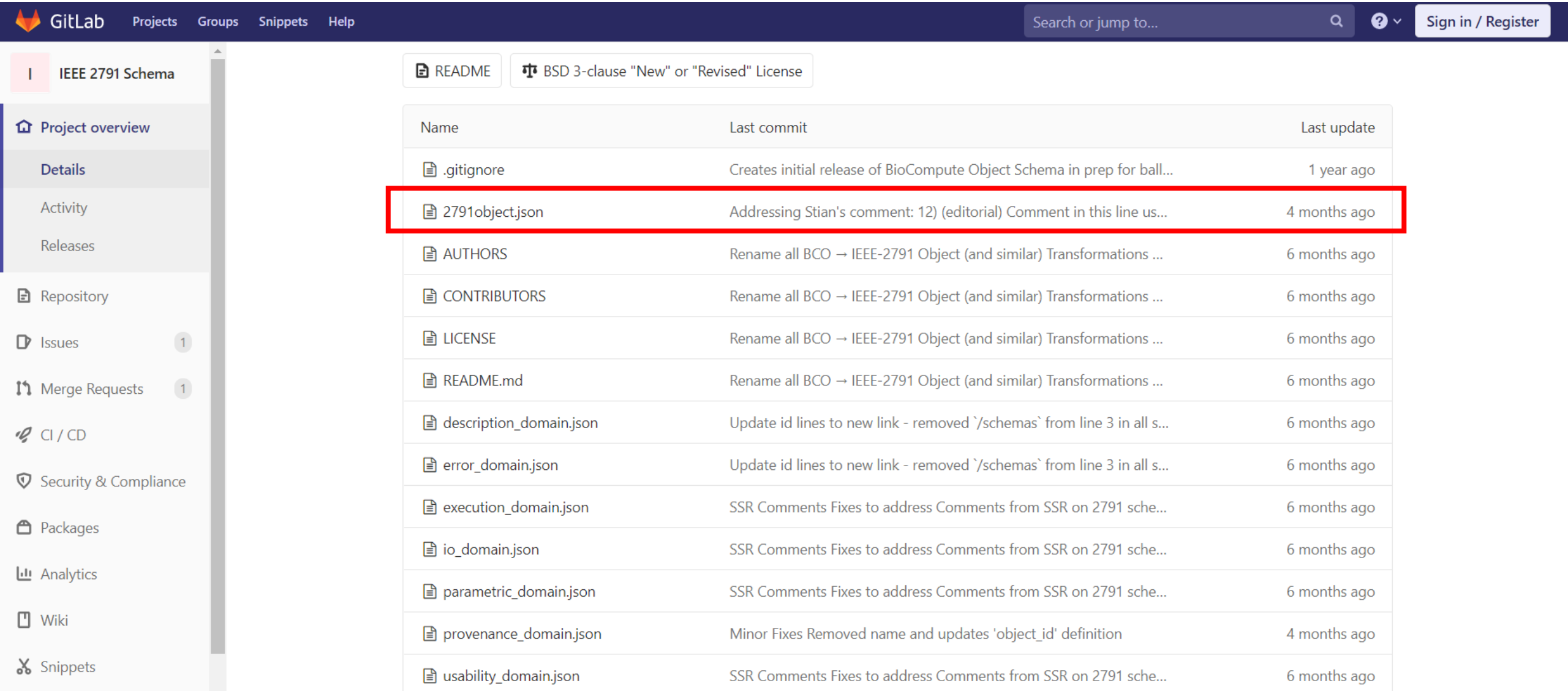
BioCompute Schema Files

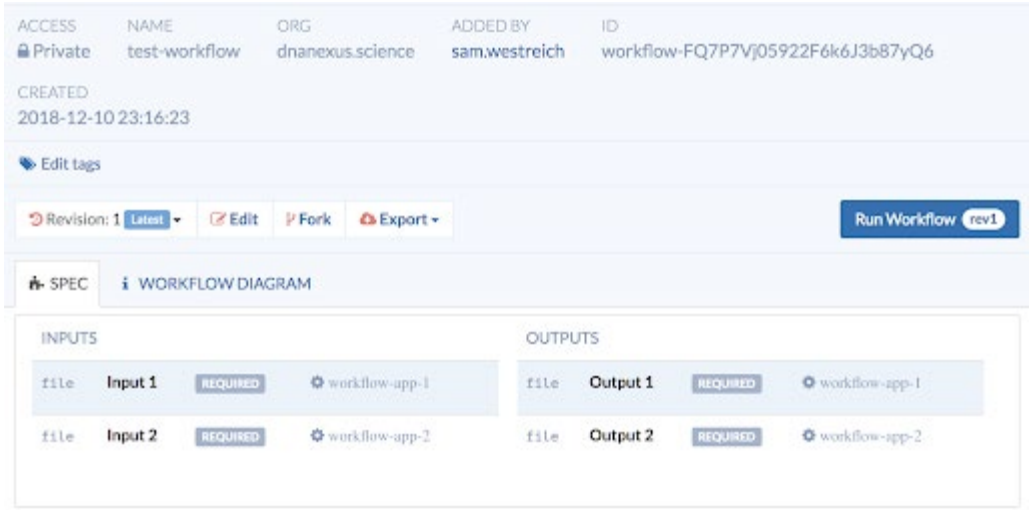
https://gitlab.com/IEEE-SA/2791/ieee-2791-schema



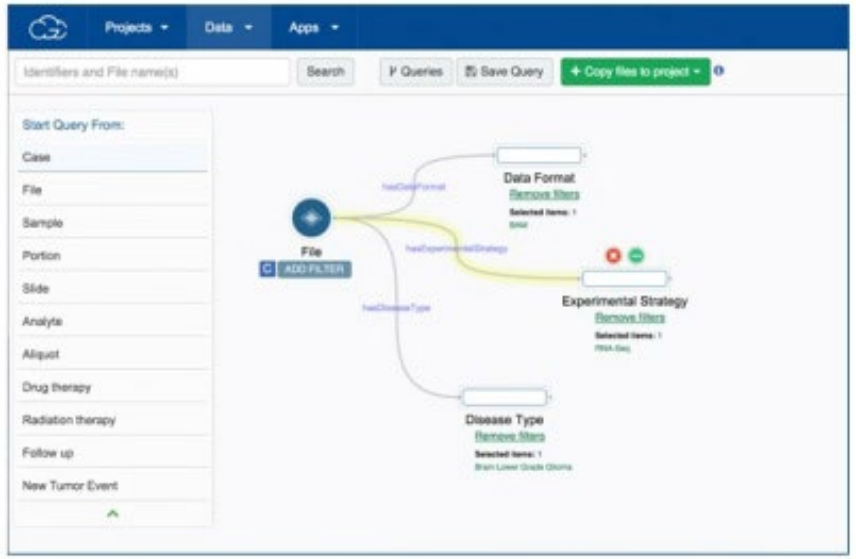
BioCompute Schema Files

https://gitlab.com/IEEE-SA/2791/ieee-2791-schema



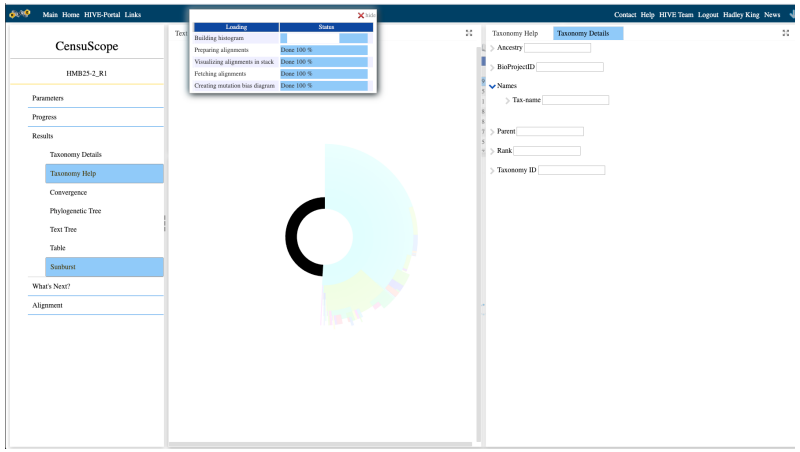


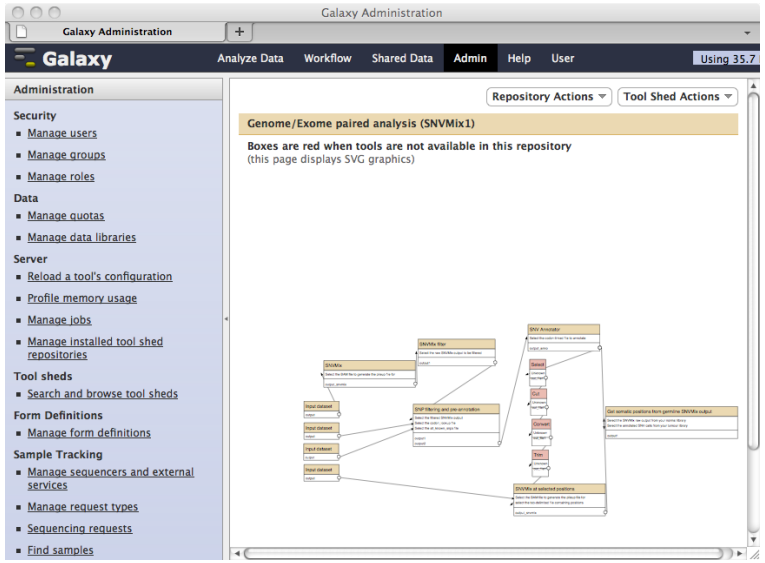
Seven Bridges



precisionFDA

DNAnexus
powering the genomics revolution™



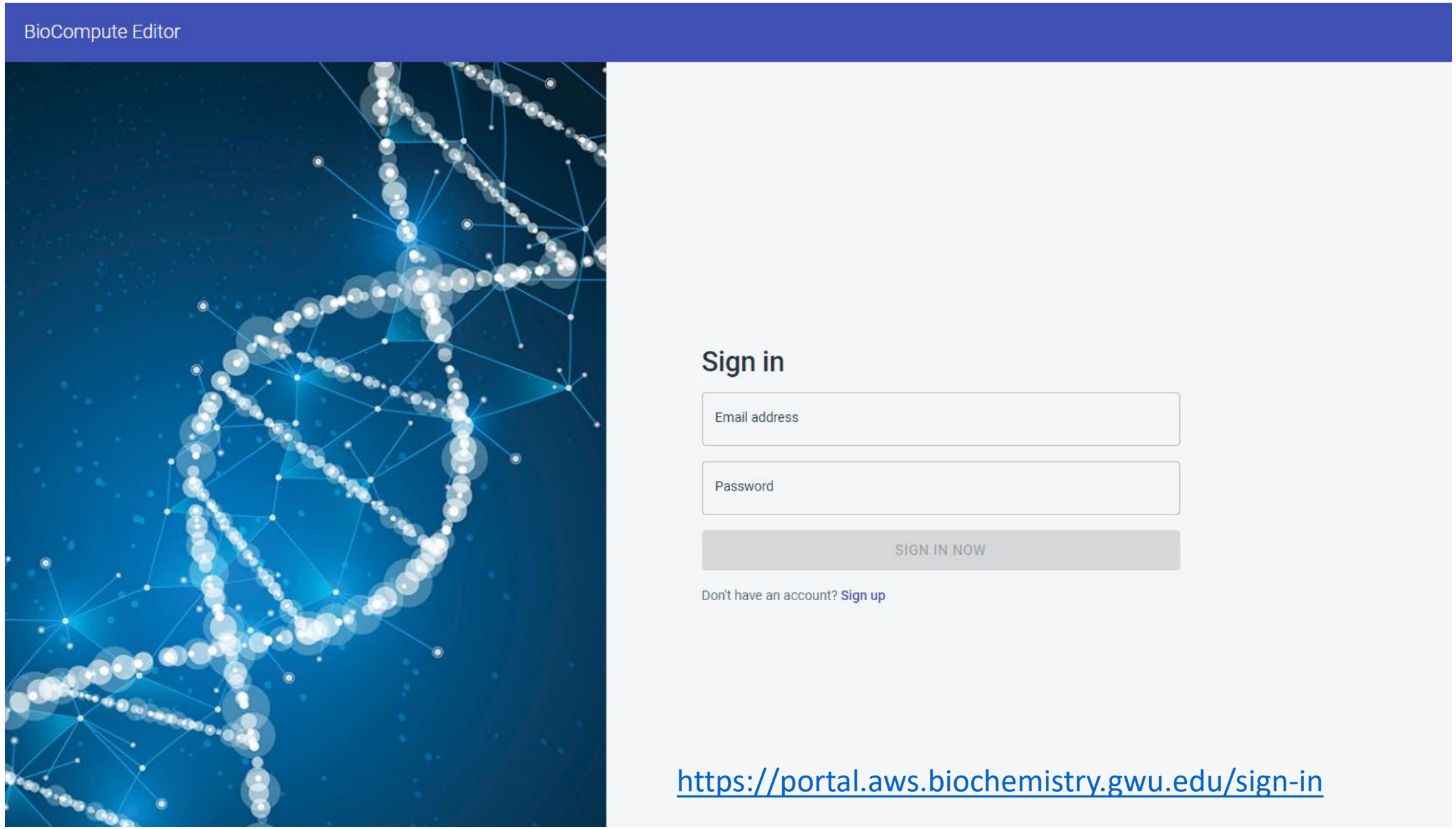


Galaxy PROJECT

HIVE

BioCompute Objects

BCO Portal



https://portal.aws.biochemistry.gwu.edu/sign-in

Children's National
precisionFDA
ATTAIN
Global Alliance for Genomics & Health
DISEASE ONTOLOGY
EPINOMICS
NIH NATIONAL CANCER INSTITUTE
CUROVERSE
Forum for Collaborative HIV Research
Google
CDC CENTERS FOR DISEASE CONTROL AND PREVENTION
EMBL-EBI
DNAnexus
HARVARD UNIVERSITY
NATIONAL LIBRARY OF MEDICINE
NCBI
FHIR
FHIR Genomics
paradigm4
Oxford e-Research Centre
UNIVERSITY OF OXFORD
Seven Bridges
NIST
Galaxy PROJECT
NEW YORK GENOME CENTER
GENOME NEXT
CRITICAL PATH INSTITUTE
NIH
research object.org
HL7
CWL
FLATIRON
MANCHESTER 1824
The University of Manchester
elixir UNITED KINGDOM
CUNY

Use-Case Examples

Tuberculosis Detection

- Tuberculosis (TB) is top infectious killer in the world
- WHO is adopting ReSeqTB pipeline to address the many challenges of detecting TB
- Requires lineage identification, prediction of antibiotic resistance, recurrence of TB in previously treated patients

Test Submission

- HCV-1a use case using synthesized data
- What data are necessary to make a regulatory decision?
- Are summary data from one analysis pipeline sufficient?
- How will the analysis pipeline be validated?

Embleema

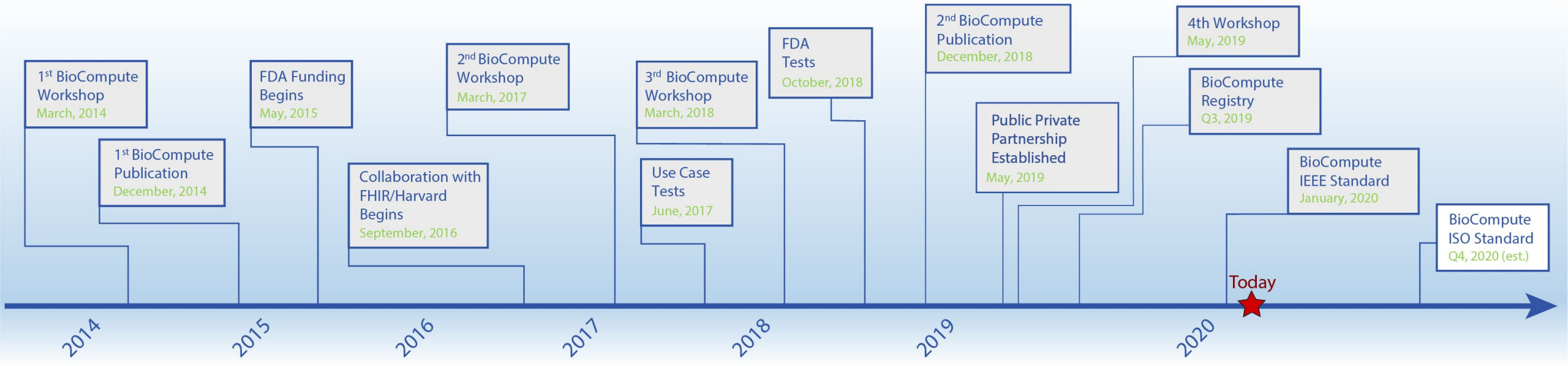
- Embleema is a platform that allows users to take control of their own data
- Marketplace for directly selling personal genome data
- Aggregator for Real World Evidence

Integrating with Other Standards

- Institute of Electrical and Electronics Engineers Standard
  - Approved January 2020; anticipated publication on April 10, 2020
- International Standards Organization certification expected by Q4 2020 through joint agreement

BioCompute Objects

BCO Timeline



BioCompute Objects

* Complete conformance check details are available on challenge site

BioCompute Object (BCO) App-a-thon

May 14 through October 18

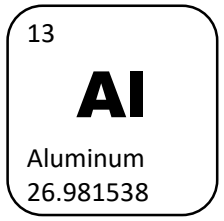
	Beginner Track: Create BCOs	Advanced Track: Create BCO software tools	
BCOs and Tools	Complete BCO	Conformance Tool Writer Tool	**May 14 – October 18, 2019** Participants create BCOs and BCO tools
Conformance Check*	Ability of an FDA reviewer to understand the BCO	Ability to generate accurate BCO; Ability to conduct conformance check on BCOs	**Deadline: October 18, 2019** Evaluation begins
Results			

App-a-thon Launch: May 14, 2019
Challenge details are made available

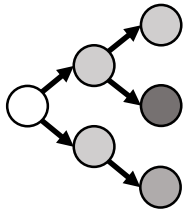
HIVE precisionFDA THE GEORGE WASHINGTON UNIVERSITY WASHINGTON, DC

Modeling Simulation

- BioGaming
  - Similar to “war gaming”
  - Data driven simulations of a pandemic
  - Provides insights into how interventions or pathogen changes may affect pandemic course



- Aluminum Efflux
  - CBER algorithm for creating aluminum pharmacokinetic profiles
    - Mitkus et al., 2011
  - Safety prediction of aluminum containing infant vaccines
  - Predicts Aluminum safety threshold
    - Based on level set by the Agency for Toxic Substances and Disease Registry



- nSARS-CoV-2 Evolution Tracking
  - CDER-based pipeline for identifying variants
  - Also used to inform glycospike protein

https://www.ncbi.nlm.nih.gov/pubmed/?term=Vaccine+29%3A+9538-9543

BioCompute Objects

Contact

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

BioCompute Objects