



BioCompute – A Framework for Community Harmonization of High-Throughput Sequencing Computational Pipelines

Elaine E. Thompson¹, Vahan Simonyan¹, Raja Mazumder² and the HIVE Team^{1,2}

¹Food and Drug Administration, Center for Biologics Evaluation and Research, Silver Spring, MD, 20993

²Department of Biochemistry and Molecular Biology, The George Washington University, Washington, DC, 20037

THE GEORGE
WASHINGTON
UNIVERSITY
WASHINGTON, DC

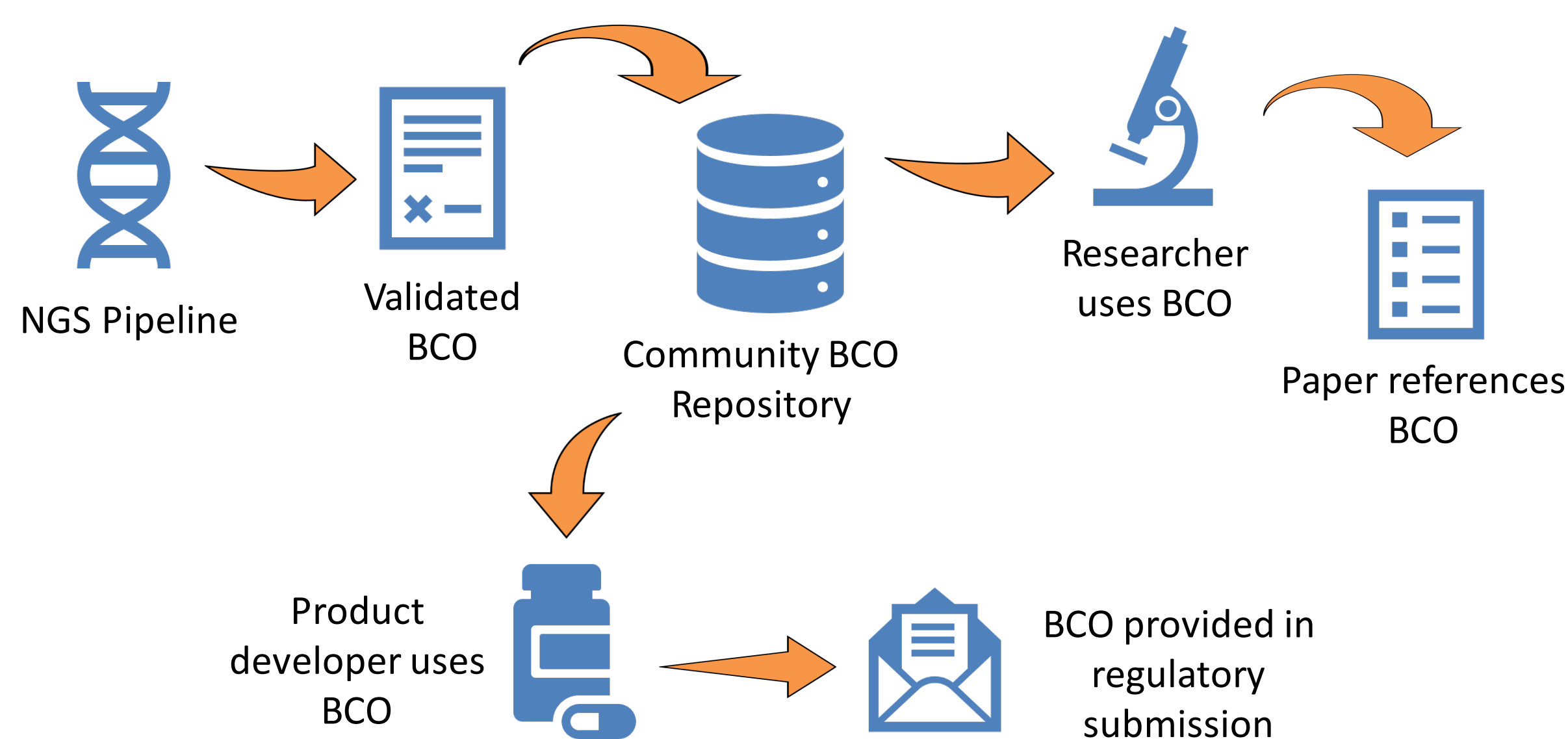
ABSTRACT

The BioCompute Framework provides a structured framework for communication of High-throughput sequencing (HTS) analysis and results. It is standard practice to provide experimental conditions when reporting scientific results; however, equivalent information on parameters and software used to generate computational results is often missing or incomplete. Additionally, data files may be nonstandard, proprietary, or otherwise difficult to use. BioCompute addresses the issue through community based standards development and harmonization of HTS computations and data formats, promoting interoperability, supporting reproducibility and simplifying the review of bioinformatics pipelines and results. A BioCompute Object (BCO) includes information about software and settings, test data, the purpose of the pipeline, references, keywords, and authors. The BCO can be provided to a collaborator or included in a regulatory submission and viewed or executed in a genomics workflow platform like the FDA HIVE. The BioCompute Framework serves to ensure interoperability between platforms, industries, scientists, regulators, and researchers.

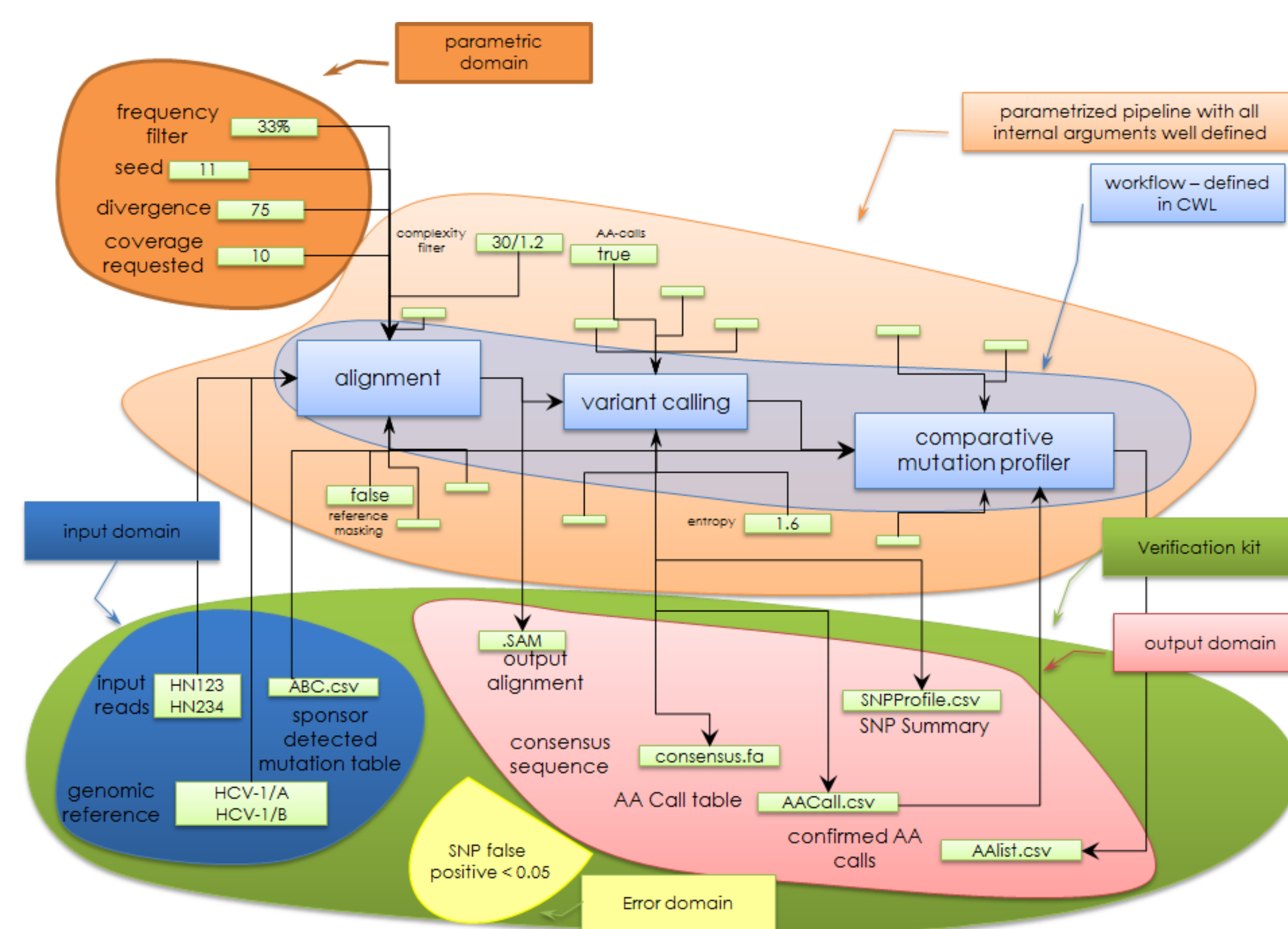
USES FOR BIOCOMPUTE OBJECTS

- A pharmaceutical company is submitting NGS data and the FDA conducts a reanalysis of the data. The reanalysis does not concur with the original results. It can be very lengthy and costly to figure out where the discrepancies lie. Attaching a biocompute object with the initial submission would simplify identification of the discrepancies.
- A research scientist performs computations and submits to a journal for peer review. The reviewers determine the results are not documented well enough to be reproducible, perhaps due to the fact that not all the details are present. Submission of a complete BioCompute object with the publication would meet the standard scientific practice of providing fully detailed methodology in order to allow peer review and reproduction of the research.
- A regulatory decision has been made where a computational analysis has been used as evidence. New data emerges after the product has been on the market over a year and the regulators can not reproduce the original environment without detailed specifications of software and parameters used in the analysis pipeline. Providing a BioCompute object would allow the computation to be reproduced at a later time.
- Authors and pharmaceutical scientists are interested in how regulatory agencies are using workflows to analyze data. Openness and transparency are hindered by the lack of ability to communicate, not a lack of willingness. BioCompute provides a common “language” for communicating computations.

BIOCOMPUTE COMMUNITY SHARING



BIOCOMPUTE SCHEMATIC



EXAMPLE BIOCOMPUTE JSON FOR ANTIVIRAL RESISTANCE DETECTION

```
{
  "id": "obj.1270",
  "type": "antiviral_resistance_detection",
  "name": "HCV1a [taxID:31646] ledipasvir [PubChem:67505836] resistance SNP [SO:0000694] detection",
  "version": "1.1",
  "digital_signature": "905d7fce3f3ac64c8ea86f058ca71658",
  "verification_status": "unreviewed",
  "publication_status": "draft",
  "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", "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",
    "authors": [{"name": "Eric Donaldson"}, {"orcid": "0000-0003-1409-4549"}],
    "description_domain": {
      "xref": [{"SO:0000694", "SO:0000045", "PubChem:67505836", "taxID:31646", "PMID: 25123381", "PMID: 26508693"}],
      "keywords": ["antiviral resistance", "SNP"],
      "pipeline_steps": {"HIVE_hexagon": {...}, "HIVE_heptagon": {...}},
      "execution_domain": {
        "platform": "hive",
        "pipeline_version": "0.1",
        "driver": "shell",
        "url": "https://hive.biochemistry.gwu.edu/workflows/",
        "env_parameters": [{"HIVEv1.3"}],
        "script": "hive://workflows/antiviral_resistance.py",
        "prerequisites": [
          {"name": "HIVE_hexagon", "version": "1.3"}, {"name": "HIVE_heptagon", "version": "1.3"}],
        "parametric_domain": {
          "hexagon_minimum_coverage": "0.15",
          "hexagon_seed": "14",
          "hexagon_minimum_match_len": "66",
          "heptagon_freq_cutoff": "0.10",
          "heptagon_divergence_threshold_percent": "30",
          "io_domain": {
            "reference_uri": ["http://www.ncbi.nlm.nih.gov/nuccore/CP000139.1"],
            "input_uri_list": ["hive://nuc-read/514683", "hive://nuc-read/514682"],
            "output_uri_list": ["hive://data/514769/dnaAccessionBased.csv", "hive://data/514801/SNP.csv"],
            "error_domain": [{"false negative discovery < 0.10", "false positive discovery < 0.05"}]
          }
        }
      }
    }
  ]
}
```

INTEROPERABILITY AND SPECIFICATION

The BioCompute framework follows Findability, Accessibility, Interoperability, and Reusability (FAIR) principles and is compatible with Common Workflow Language (CWL) and HL7 FHIR.

The BioCompute Object Project and specification are available at OSF. <https://osf.io/h59uh/>

Disclaimer: This poster is an informal communication and represents the authors’ best judgment. This communication does not bind or obligate FDA.

ACKNOWLEDGEMENTS

Mark Walderhaug, Carolyn Wilson, **HIVE team**: Grace Alterovitz, Amanda Bell, Prince Birring, Ting-Chia Chang, Stefan Dabic, Hiral Desai, Hayley Dingerdisen, Lama Elzohary, Marianna Faradzheva, Brian Fochtman, Tigran Ghazanchyan, Anton Golikov, Naila Gulzar, Robel Kahsay, Konstantinos Karagiannis, Charles Hadley King, Ilya Mazo, Reza Mousavi, Rahi Navelkar, Ekaterina Osipova, Aleksey Pshenichnov, Yao Ren, Alexandre Rostovtsev, Xutian Ruan, Luis Santana-Quintero, Michelle Shen, Amanraj Singh, Krista Smith, Valery Tkachenko, Phuc VinhNguyen Lam, Jeet Kiran Vora, Alin Voskanian