Enabling Precision Medicine via standard communication of NGS provenance, analysis, and results

Gil Alterovitz[1,2,3], Dennis Dean[4], Carole Goble[5], Michael R. Crusoe[6], Stian Soiland-Reyes[5], Amanda Bell[7,8], Anais Hayes[7,8], Anita Suresh[24], Charles H. King[7,8], Dan Taylor[9], KanakaDurga Addepalli[13,14], Elaine Johanson[10], Elaine E. Thompson[10], Eric Donaldson[10], Hiroki Morizono[11,12], Hsinyi Tsang[13,14], Jeet K. Vora[7,8], Jeremy Goecks[15], Jianchao Yao[16], Jonas S. Almeida[17], Konstantinos Krampis[18,19], Krista M. Smith[10], Lydia Guo[20], Mark Walderhaug[10], Marco Schito[25], Matthew Ezewudo[25], Nuria Guimera[23], Paul Walsh[21], Robel Kahsay [7,8] Srikanth Gottipati[26], Timothy C Rodwell[24], Toby Bloom[22], Yuching Lai[23], Vahan Simonyan[10*], Raja Mazumder[7,8*]

[1] Harvard/MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA
[2] Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115, USA
[3] Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Boston, MA 02139, USA,
[4] Seven Bridges, Cambridge MA, 02142, USA
[5] School of Computer Science, The University of Manchester, Manchester, M13 9PL, UK
[6] Common Workflow Language Project, Vilnius, Lithuania
[7] The Department of Biochemistry & Molecular Medicine, The George Washington University Medical Center, Washington, DC 20037, USA
[8] The McCormick Genomic and Proteomic Center, The George Washington University, Washington, DC 20037, USA
[9] Internet 2, 1150 18th St. NW, Washington, DC 20036, USA
[10] US Food and Drug Administration, Silver Spring MD 20993, United States of America
[11] Center for Genetic Medicine, Children's National Medical Center, Washington, DC 20010, USA
[12] The Department of Genomics and Precision Medicine, The George Washington University School of Medicine and Health Sciences, Washington, DC 20037, USA
[13] Center for Biomedical Informatics and Information Technology, National Cancer Institute, National Institutes of Health, Gaithersburg, MD, USA
[14] Attain, LLC, McClean, VA, USA
[15] Computational Biology Program, Oregon Health & Science University, Portland OR, 97239, USA
[16] MRL IT, Merck & Co., Inc., Boston, MA, USA
[17] Stony Brook University, School of Medicine and College of Engineering and Applied Sciences, Stony Brook, NY 11794, USA
[18] Department of Biological Sciences, Hunter College of The City University of New York, USA
[19] Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY 10021, USA
[20] Wellesley College, Wellesley, MA 02481, USA
[21] NSilico Life Science, Nova Center, Belfield Innovation Park, University College Dublin, Dublin 4, Ireland
[22] New York Genome Center, New York, NY 10013, USA
[23] DDL Diagnostic Laboratory, 2288 ER, Rijswijk, Netherlands
[24] Foundation for Innovative New Diagnostics (FIND), Chemin des Mines 9, 1202 Geneva, Switzerland
[25] Critical Path Institute, Tucson, AZ, USA
[26] OTSUKA Pharmaceutical Development & Commercialization, Inc, Princeton, NJ, USA

*Corresponding authors

Abstract. Precision medicine can be empowered by a personalized approach to patient care based on the patient's or pathogen's unique genomic sequence. For precision medicine, genomic findings must be robust and reproducible, and experimental data capture should adhere to FAIR guiding principles. Moreover, precision medicine requires standardized reporting that extends beyond wet lab procedures to computational methods. Rapidly developing standardization technologies can improve communication and reporting of genomic sequence data analysis steps by utilizing concepts defined in the BioCompute framework, such as error domain, usability domain, verification kit, and provenance domain. These advancements allow data provenance to be standardized and promote interoperability. Thus, a resulting bioinformatics computation instance that includes these advancements can be easily communicated, repeated and compared by scientists, regulators, test developers and clinicians. Easing the burden of doing the aforementioned tasks greatly extends the range of practical application. Advancing clinical trials, precision medicine, and regulatory submissions requires an umbrella of standards that not only fuses these elements, but also ensures efficient communication and documentation of genomic analyses. The BioCompute paradigm and the resulting BioCompute Objects (BCOs) offer that umbrella. Through standardized bundling of high-throughput sequencing studies under BCOs, regulatory agencies (e.g., FDA), test developers, researchers, and clinicians can expand collaboration to drive innovation in precision medicine, with the potential for decreasing the time and cost associated with next generation sequencing workflow exchange, reporting, and regulatory reviews.

Keywords: BioCompute, BioCompute Objects, high-throughput sequencing (HTS), next generation sequencing (NGS), regulatory review, CWL, FHIR, GAG4H, HL7, research objects, provenance, FAIR data guidelines.

Introduction

Precision medicine practice must capture the process of producing, sharing, and consuming genomics information. Capturing the process of genomic data generation will empower individuals, research institutes, clinical organizations, and regulatory agencies to evaluate and trust the reliability of biomarkers

generated from complex analyses (e.g., presence of a specific variant). Efforts to promote data-sharing structures for genome-wide association studies (GWAS) offer additional benefits, as seen in the National Center for Biotechnology Information's (NCBI) Database of Genotypes and Phenotypes (dbGaP)[1] and ClinVar[2] as well as LD Hub, a centralized database of GWAS results for diseases/traits[3]. Although, the importance of data sharing is well accepted, discussions related to recording, reporting, and sharing of analysis protocols are often overlooked.

The need for NGS provenance, analysis, and results is critical as we enter the clinical genomic era. The cost for High Throughput Sequencing (HTS) has fallen from $20 per base in 1990 to less than $.01 per base in 2011, creating massive data accumulation[4]. Lower costs of HTS data generation have increased the availability of data, expediting more types of analyses. In recent years, there has been a focus on novel drug development and precision medicine research to create innovative, reliable, and accurate -*omics*-based (i.e., genomics, transcriptomics, proteomics) tests[5]. These initiatives allow different data sources to be analyzed by a variety of methods advancing genomic analyses. Often, information about the bioinformatics pipelines used in these analyses is not reported in a format that is easily understood, comprehensive and interoperable. Without a universal standard for communicating how these data are analyzed to obtain a specific result, we quickly encounter the "tower-of-Babel problem" – a problem of diversified languages and miscommunication.

Currently, there are some standards to capture genomic sequencing information and provenance. Fast Healthcare Interoperability Resources[6,7] (FHIR) and organizations such as Global Alliance for Genomics and Health[8] (GA4GH) communicate genomic information, although these cater to specific community domains. The Common Workflow Language[9] (CWL) and Research Objects[10] (RO) capture reproducible workflows in a domain agnostic manner. The BioCompute paradigm unites these standards to provide a provenance-reporting framework for genomic sequencing data analysis in the

context of FDA submissions and regulatory review in the form of a BioCompute Object (BCO)[11]. BCOs provide a new harmonizing approach designed to satisfy FAIR Data Principles allowing experimental data and protocols to be findable, accessible, interoperable and reusable[12] in the regulatory and research needs for evaluation, validation, and verification of bioinformatics pipelines[11,13,14]. The BCO also meets the NIH strategic plan for data science[15] which states that the quality of clinical data should be maintained at all stages of the research cycle where the BCO can be adjusted to fit the needs of a specific experiment from generation of the data through the entire analysis process.

This paper focuses on how the FHIR, GA4GH, CWL, and RO standards can be leveraged and harmonized by examining the BioCompute paradigm (See Fig. 1). Once established, the BioCompute framework can be utilized for other types of FDA submissions, such as large clinical trials, where data provenance in analysis datasets can be difficult to communicate.

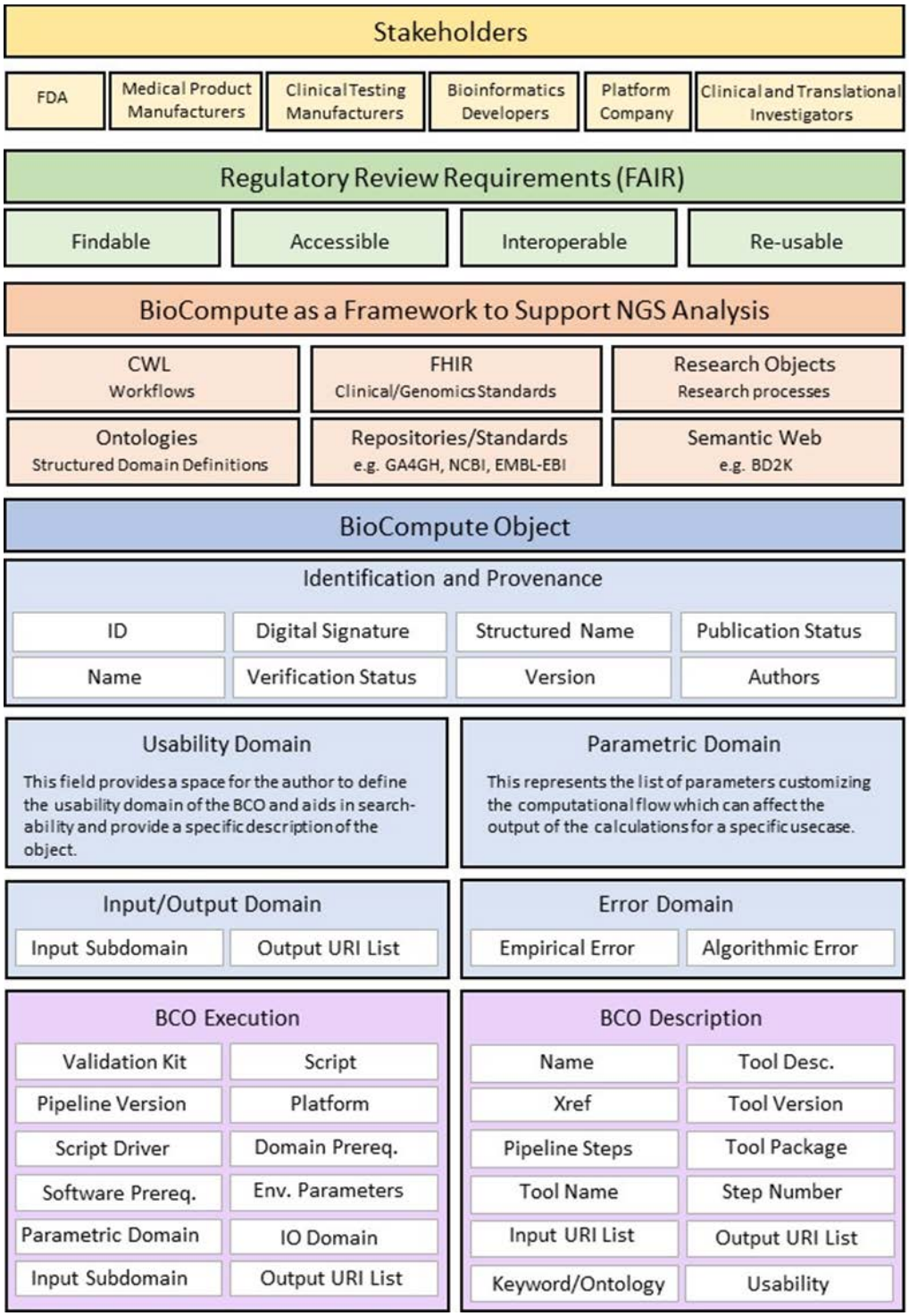

Figure 1. Schematic of BioCompute Object as a frameworks for advancing regulatory science by incorporating existing standards and introducing additional concepts that include digital signature, usability domain, validation kit, and error domain.

Background

The Academy of Medical Sciences hosted a symposium on reproducibility in pre-clinical biomedical research, where a number of measures for improving reproducibility were identified, including greater openness and transparency, defined reporting guidelines, and better utilization of standards and quality control measures[16]. Issues of reproducibility have resulted in an enormous waste of research resources and have hindered progress in the life sciences, as highlighted by several high profile articles[17,18].

Researchers propose two types of reproducibility: method reproducibility and result reproducibility[19]. The first is defined as providing detailed experimental methods so others can repeat the procedure exactly; the latter, also called replicability, is defined as achieving the same results as the original experiment by closely adhering to the methods. Method reproducibility depends significantly on the completeness of the procedures researchers provide. CWL creates a similar starting point for computational methods—if scientists gather data in the same computational language, then a large part of the provenance model is standardized. As such, CWL and provenance tracking make it easier for researchers to track errors and locate deviances. Research Objects for workflows go further, describing the packaging of the method, provenance logs, and associated data and codes with richly described metadata manifests that include the experiment's context[20].

Universally reproducible data is an outcome to aim for, but challenges remain. Without repeatability and analytics standards for NGS/HTS studies, regulatory agencies, academic researchers, pharmaceutical companies, diagnostic test developers and the FDA cannot work together to validate results and drive the emergence of new fields[21]. The lack of universally accepted standards to record a computational workflow and the subsequent analysis results in many studies that are not repeatable and therefore not usable in a clinical or practical situation. In response, various workflow management systems and

bioinformatics platforms have been developed, and each has a method to track and record computational workflows, pipelines, versions, and parameters used in the studies. However, these efforts remain haphazard and require harmonization. BCOs enable reproducible data and improved information-sharing by tracking the data's provenance and accurately documenting the trail of processes in a standard and widely applicable way.

Provenance of Data

Reproducible data requires that the data's origin and history be captured in a standardized format. Provenance here refers to a datum's history starting from the original source, namely, its *lineage*. A *lineage graph* can show the source of a datum in a database, data movement between databases or computational processes, or data generated from a computational process. Complementary to data lineage is a *process audit.* This provides a historical trail documenting a scientific study, providing snapshots of intermediary states, values for configurations and parameters, and traceability of stepwise analytical processing[22]. Such audit trails should allow an independent reviewer to effectively evaluate a computational investigation. Both types of records gather provenance information that is crucial to ensure accuracy and validity of experimental results. Modern computational workflows can now produce reams of fine-grained but not particularly useful trace records, and modern web developments make data transformation and copying easier, so gathering such material is a daunting challenge. In the molecular biology field alone, there are hundreds of public databases, but only a handful possess the "source" data; the remainder contain secondary views of the source data or views of other publications' views[23]. Accurately collecting lineage and process records, while obtaining the appropriate granularity of the record keeping and 'black-box' steps remains an ordeal[24,25].

Data-provenance tracing issues have far-reaching effects on scientific work. Reproducible experiments rely on confidence in the accuracy and validity of the data, process used, and knowledge generated (final

analysis product) especially after undergoing complex, multistep processes of aggregation, modeling, and analysis[26]. Computational investigations require interactions with adjacent disciplines and disparate fields to effectively analyze a large volume of relevant information. These hurdles require a solution beyond Open Data to establish Open Science in the community. Provenance needs to be preserved and shared to provide better transparency and reproducibility[27]. Complex analyses rely heavily on accurately shared data. Standards need to be established to communicate reliable genomic data between databases and other scientists, accurately reporting data provenance and process audit. In order to aid this, an active community has engaged in provenance standardization[28], e.g., culminating in the W3C PROV[29], used by FHIR and ROs, based on the idea of generating an entity target via an agent's activity (See Fig. 2).

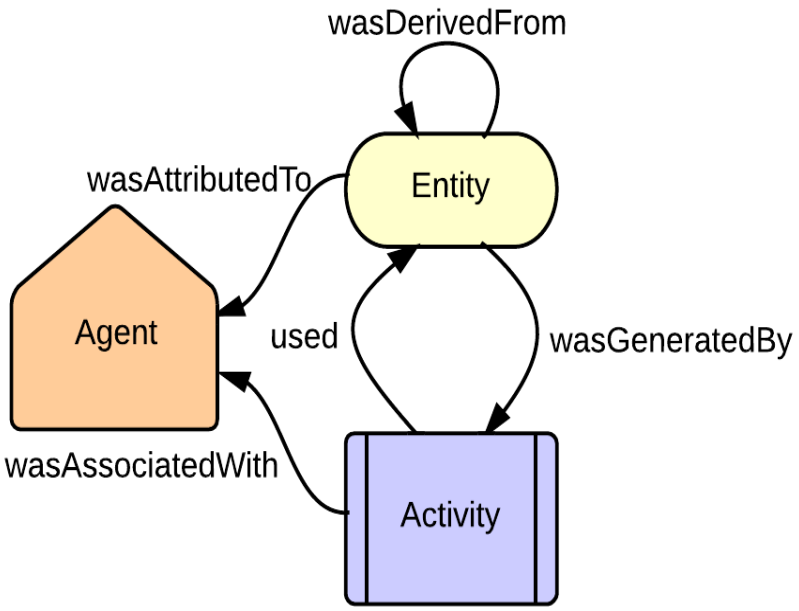

Figure 2. W3C Provenance Specification used in FHIR and ROS. Adapted from http://www.w3.org/TR/prov-primer/

Workflow management systems have been developed to capture analytic processes, and bioinformatic platforms have been built to capture analysis details. Together they can gather provenance information

for users, and extending Provenance standards where necessary. However, these systems need greater standardization. This is where BCOs can play a key role, to ensure that provenance is recreated for regulatory approval and that standards such as PROV are appropriately adopted. By properly utilizing the fields in the Identification and Provenance domain, also known as the *provenance domain,* as defined by the BioCompute framework, a BCO can become a "history of what was computed."

Key Considerations for Communication of Provenance, Analysis, and Results

Workflow Management Systems

Scientific workflows have emerged as a model to represent and manage complex distributed scientific computations[26]. Scientific workflows capture and link analysis steps and individual data transformations; each step specifies a process or computation to be executed (e.g., a software program or a web service to be invoked), and the steps are linked by the data flow and dependencies. In addition, workflows also capture the mechanisms to carry out the steps in a distributed environment, and a significant amount of work has been done to share workflows and experiments[26,30]. Workflows represent a system to capture complex analysis processes at various levels, as well as the provenance information necessary for reproducibility, result publication, and result sharing/publication [31].

Workflow management systems act to execute and monitor scientific workflows, coordinating the sequential components in a pipeline[32]. Developments in workflow management systems have led to the proposition of using workflow-centric research objects with executable components[5,20]. The use of workflow creation and management software allows researchers to utilize different resources to create complex analysis pipelines that can be executed locally, on institutional servers and on the cloud[33,34]. Extensive reviews of current workflow systems for bioinformatics are linked [33,35-37]. Ongoing systems participate in the current trend of moving from graphical systems back to script-like workflows. These systems are now executed on cloud infrastructure, high performance computing (HPC) systems, and Big

Data cluster-computation frameworks, which allow for greater data reproducibility and portability (see Supplementary Info). Workflow management systems capture provenance information, but rarely in the PROV standard. Therefore, BCOs rely on existing regulatory standards like CWL to manage pipeline details, and on ROs and FHIR to unify and enhance interoperability.

Bioinformatics platforms

The dramatic increase in the use of NGS technology for patient management has rapidly increased the need to store, access, and compute sequencing reads and other NGS/biomedical data[38]. These increased requirements have led to a call for usage methods on integrated computing infrastructure, including storage and computational nodes. This kind of integration will minimize transfer costs and remove the bottlenecks found in both downstream analyses and community communication of computational analyses results[39]. For bioinformatics platforms, communication requirements include (a) recording all analysis details such as parameters and input datasets and (b) sharing analysis details so that others can understand and reproduce analyses.

To reduce unprocessed data buildup, several high-throughput[26] cloud-based infrastructures have been developed, including HIVE (High-performance Integrated Virtual Environment)[39,40] and Galaxy[41], along with commercial platforms from companies like DNAnexus (dnanexus.com), and Seven Bridges Genomics (sevenbridges.com), among others. High throughput computing (HTC) environments deliver large amounts of processing capacity over long periods of time. These are ideal environments for long-term computation projects, as with genomic research[42]. Most HTC platforms utilize distributed cloud-computing environments to support extra-large dataset storage and computation, while hosting tools and workflows for many biological analyses. Cloud-based infrastructures also reduce the “data silo” phenomenon by converting data into reproducible formats that facilitate communication (see Supplementary Info). Additionally, the National Cancer Institute has initiated the Cloud Pilots project, in

order to test a distributed computing approach for the multi-level, large-scale datasets available on The Cancer Genome Atlas (TCGA) [43].

The genomic community has come to acknowledge the necessity of data sharing and communication to facilitate reproducibility and standardization [44]. Data sharing is crucial in everything from long-term clinical treatments to public health emergency response[45]. As industry policies develop, the need for voluntary and industry-wide standardization becomes undeniable. Extending bioinformatics platforms to include data provenance, standard workflow computation, and encoding results with available standards through BCO implementation will greatly support the exchange of genomic data analysis methods for regulatory review.

Regulatory Supporting Standards

Assessment of data submitted in a regulatory application requires clear communication of data provenance, computational workflows, and traceability. A reviewer must be able to verify that sequencing was done appropriately, pipelines and parameters were applied correctly, and that the final result, like an allelic difference or variant call, is valid. Because of these requirements, review of any clinical trial or any submission supported with NGS results requires considerable time and expertise. Submission of a BCO would ensure that data provenance is unambiguous and that the bioinformatics workflow is fully documented[11,15,23,46,47].

To truly understand and compare computational tests, a standard method (like BCO) requires tools to capture progress and to communicate the workflow and input/output data. As the regulatory field progresses, methods have been developed and are being continually refined to capture workflows and exchange data electronically[26]. See Figure 3 for BCO extensions to NGS analysis to support data provenance and reproducibility.

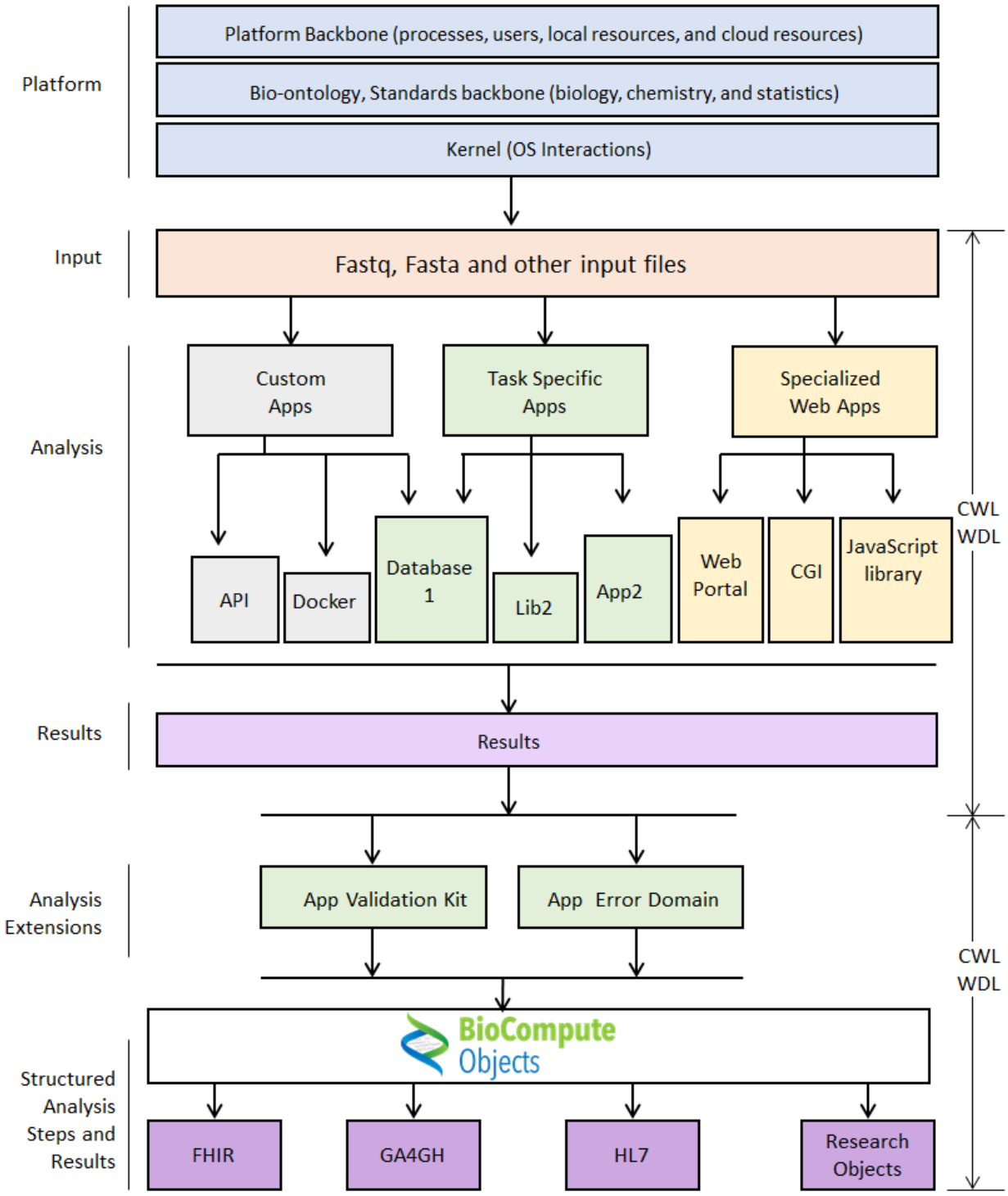

Figure 3. Generic NGS platform schematic with proposed BioCompute Object integrations and extensions.

Biocompute Objects (BCOs) and Their Harmonizing Efforts

Biocompute Objects (BCOs) were conceptualized to alleviate the disparate nature of NGS computational analyses. The primary objectives of BCOs are to (a) harmonize NGS computational results and data formats and (b) encourage interoperability and success in the verification of bioinformatics protocols. Harmonizing the above standards is especially applicable to clarify genomics/workflow instance provenance for FDA submissions. Each BCO comprises information on the arguments and versions of executable programs in a pipeline, references to input/output data, a usability domain, keywords, a list of authors, and other important sources of metadata. The BCO can serve as a wrapper. Enabling BCOs to incorporate existing standards provides a universal framework for including existing advances in workflow and data specifications. What is novel about the BioCompute paradigm is the combination of existing standards with the methodologies and tools to evaluate an experiment both programmatically and empirically. The BCO takes a snapshot of an experiment's entire computational procedure adhering to FAIR data guidelines by making it findable through the BCO portal, accessible, interoperable and maintaining it's richness to make it reusable. Using this snapshot of the BCO, which includes the results from the experiment in the dataset (verification kit), allows any other user to run the exact experiment and produce the same results. The verification kit also allows a BCO to be assessed once someone has decided that they want to change parameters, for example, and use the same BCO. Additionally, through the use of provenance and usability domains, a knowledgeable reviewer can quickly decide if the underlying scientific principles merit approval or further review.

Discussion and Conclusions

Robust and reproducible data analysis is key to successful personalized medicine and genomic initiatives. Researchers, clinicians, administrators, and patients are all tied by the information in electronic health records (EHRs) and databases. Current systems rely on data stored with incomplete

provenance records and different computing languages. This has created a cumbersome and inefficient healthcare environment.

The initiatives discussed in this review seek to make data and analyses robust and reproducible to facilitate collaboration and information sharing from data producers to data users. Increased NGS/HTS sequencing creates silos of unusable data, making standardized regulation of reproducibility more difficult. To clear the bottleneck of downstream analysis, the provenance (or origin) of data along with the analysis details (e.g., parameters, workflow versions), must be tracked to ensure accuracy and validity. The development of high-throughput cloud-based infrastructures like DNAnexus, Galaxy, HIVE, and Seven Bridges Genomics enables users to capture data provenance and store the analyses in infrastructures that allow easy user interaction.

Platform-independent provenance has largely been ignored in HTS. Emerging standards enable both representation of genomic information and linking of provenance information. By harmonizing across these standards, provenance information can be captured across both clinical and research settings extending into the conceptual experimental methods and the underlying computational workflows. There are several use cases of such work, including submission for FDA diagnostic evaluations, as is being done with the BCO effort. Such standards also enable robust and reproducible science, and facilitate open science between collaborators. At this juncture, it is imperative to lead the development of these standards to satisfy the needs of downstream consumers of genomic information.

The need to reproducibly communicate HTS/NGS computational analyses has led to collaboration among disparate industry players. Through various conferences/workshops, attention has increased exposure to standardization, tracking, and reproducibility methods[48,49]. Standards like FHIR and ROs capture the underlying data provenance to be shared in frameworks like GA4GH, enabling collaboration

around reproducible data and analyses. New computing standards like Common Workflow Language (CWL) increase the scalability and reproducibility of data analysis. The BioCompute paradigm acts as a harmonizing umbrella to facilitate data submitted to regulatory agencies, increasing interoperability in the genomic field. BioCompute specifications, available at https://osf.io/h59uh/, can be used to generate BCOs by any bioinformatics platform that automatically pulls underlying data and analysis provenance into its infrastructure. Ongoing BCO pilots are currently working to streamline the flow to provide users with effortlessly reproducible bioinformatics analyses. As BCOs aim to simplify FDA approval, these pilots mirror clinical trials involving NGS data for FDA submissions. Fusing bioinformatics platforms and HTS standards to capture data and analyze provenance for BCOs make robust and reproducible analyses and results an attainable standard for the scientific community.

Acknowledgements

We would like to recognize all the speakers and participants of the 2017 HTS-CSRS workshop who facilitated the discussion on standardizing HTS computations and analyses. The workshop was co-sponsored by FDA and GW. The comments and input during the workshop were processed and integrated into the BCO specification document and this manuscript. The participants of the 2017 HTS-CSRS workshop are listed here: https://osf.io/h59uh/wiki/2017%20HTS-CSRS%20Workshop/. This work has been funded in part by FDA (HHSF223201510129C) and The McCormick Genomic and Proteomic Center at the George Washington University.

Disclaimer

The contributions of the authors are an informal communication and represent their own views.

References

1. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, et al. (2007) The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39: 1181-1186.

2. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, et al. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42: D980-985.

3. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, et al. (2017) LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics 33: 272-279.

4. Sawyer E (2017) High Throughput Sequencing and Cost Trends. Nature Education.

5. (2012) In: Micheel CM, Nass SJ, Omenn GS, editors. Evolution of Translational Omics: Lessons Learned and the Path Forward. Washington (DC).

6. Beredimas N, Kilintzis V, Chouvarda I, Maglaveras N (2015) A reusable ontology for primitive and complex HL7 FHIR data types. Conf Proc IEEE Eng Med Biol Soc 2015: 2547-2550.

7. Alterovitz G, Warner J, Zhang P, Chen Y, Ullman-Cullere M, et al. (2015) SMART on FHIR Genomics: facilitating standardized clinico-genomic apps. J Am Med Inform Assoc 22: 1173-1178.

8. Lawler M, Siu LL, Rehm HL, Chanock SJ, Alterovitz G, et al. (2015) All the World's a Stage: Facilitating Discovery Science and Improved Cancer Care through the Global Alliance for Genomics and Health. Cancer Discov 5: 1133-1136.

9. Peter Amstutz MRC, Nebojša Tijanić (editors), Brad Chapman, John Chilton, Michael Heuer, Andrey Kartashov, Dan Leehr, Hervé Ménager, Maya Nedeljkovich, Matt Scales, Stian Soiland-Reyes, Luka Stojanovic (2016) Common Workflow Language,. Specification, Common Workflow Language working group. .

10. Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, et al. (2013) Why linked data is not enough for scientists. Future Generation Computer Systems-the International Journal of Grid Computing and Escience 29: 599-611.

11. Simonyan V, Goecks J, Mazumder R (2017) Biocompute Objects-A Step towards Evaluation and Validation of Biomedical Scientific Computations. PDA J Pharm Sci Technol 71: 136-146.

12. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3: 160018.

13. Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118: 1590-1605.

14. Boyd SD (2013) Diagnostic applications of high-throughput DNA sequencing. Annu Rev Pathol 8: 381-410.

15. NIH (2018) NIH STRATEGIC PLAN FOR DATA SCIENCE. In: Research OoE, editor.

16. Bishop D (2015) Reproducibility and reliability of biomedical research. The Academy of Medical Sciences.

17. Pusztai L, Hatzis C, Andre F (2013) Reproducibility of research and preclinical validation: problems and solutions. Nat Rev Clin Oncol 10: 720-724.

18. Samuel Reich E (2011) Cancer trial errors revealed. Nature 469: 139-140.

19. Goodman SN, Fanelli D, Ioannidis JP (2016) What does research reproducibility mean? Sci Transl Med 8: 341ps312.

20. Belhajjame K, Zhao J, Garijo D, Gamble M, Hettne K, et al. (2015) Using a suite of ontologies for preserving workflow-centric research objects. Journal of Web Semantics 32: 16-42.

21. Kjer KM, Gillespie JJ, Ober KA (2007) Opinions on multiple sequence alignment, and an empirical comparison of repeatability and accuracy between POY and structural alignment. Syst Biol 56: 133-146.

22. Bose R, Frew J (2005) Lineage retrieval for scientific data processing: A survey. Acm Computing Surveys 37: 1-28.

23. Buneman P, Khanna, S. & Wang-Chiew, T (2001) Why and Where: A Characterization of Data Provenance. In Database Theory. Springer Lecture Notes in Computer Science: pp. 87–93.

24. Freire J, Bonnet, P. & Shasha, D. (2012) Computational Reproducibility: State-of-the-art, Challenges, and Database Research Opportunities. SIGMOD Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data: pp. 593–596.

25. Alper P. Enhancing and Abstracting Scientific Workflow Provenance for Data Publishing; 2013.

26. Gil Y, Deelman E, Ellisman M, Fahringer TF, Fox G, et al. (2007) Examining the challenges of scientific workflows. Computer 40: 24-+.

27. Reichman OJ, Jones MB, Schildhauer MP (2011) Challenges and Opportunities of Open Data in Ecology. Science 331: 703-705.

28. Moreau L, Clifford B, Freire J, Futrelle J, Gil Y, et al. (2011) The Open Provenance Model core specification (v1.1). Future Generation Computer Systems-the International Journal of Grid Computing and Escience 27: 743-756.

29. Ciccarese P, Soiland-Reyes S, Belhajjame K, Gray AJ, Goble C, et al. (2013) PAV ontology: provenance, authoring and versioning. J Biomed Semantics 4: 37.

30. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, et al. (2010) myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic Acids Res 38: W677-682.

31. Garijo D, Gil Y, Corcho O (2017) Abstract, link, publish, exploit: An end to end framework for workflow sharing. Future Generation Computer Systems-the International Journal of Escience 75: 271-283.

32. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, et al. (2010) myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic Acids Res 38: W677-W682.

33. Cohen-Boulakia S, Belhajjame K, Collin O, Chopard J, Froidevaux C, et al. (2017) Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. Future Generation Computer Systems-the International Journal of Escience 75: 284-298.

34. Woodcock J, Woosley R (2008) The FDA critical path initiative and its influence on new drug development. Annu Rev Med 59: 1-12.

35. Leipzig J (2017) A review of bioinformatic pipeline frameworks. Brief Bioinform 18: 530-536.

36. Spjuth O, Bongcam-Rudloff E, Hernandez GC, Forer L, Giovacchini M, et al. (2015) Experiences with workflows for automating data-intensive bioinformatics. Biol Direct 10: 43.

37. Xu J, Thakkar S, Gong B, Tong W (2016) The FDA's Experience with Emerging Genomics Technologies-Past, Present, and Future. AAPS J 18: 814-818.

38. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46.

39. Simonyan V, Mazumder R (2014) High-Performance Integrated Virtual Environment (HIVE) Tools and Applications for Big Data Analysis. Genes (Basel) 5: 957-981.

40. Simonyan V, Chumakov K, Dingerdissen H, Faison W, Goldweber S, et al. (2016) High-performance integrated virtual environment (HIVE): a robust infrastructure for next-generation sequence data analysis. Database (Oxford) 2016.

41. Afgan E, Baker D, Coraor N, Goto H, Paul IM, et al. (2011) Harnessing cloud computing with Galaxy Cloud. Nat Biotechnol 29: 972-974.

42. Thain D, Tannenbaum T, Livny M (2005) Distributed computing in practice: the Condor experience. Concurrency and Computation-Practice & Experience 17: 323-356.

43. Tomczak K, Czerwinska P, Wiznerowicz M (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn) 19: A68-77.

44. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P (2009) Data sharing in genomics--re-shaping scientific practice. Nat Rev Genet 10: 331-335.

45. Whitty CJ (2017) The contribution of biological, mathematical, clinical, engineering and social sciences to combatting the West African Ebola epidemic. Philos Trans R Soc Lond B Biol Sci 372.

46. Kenall A, Harold S, Foote C (2014) An open future for ecological and evolutionary data? BMC Evol Biol 14: 66.

47. Buneman P, Khanna, S. & Tan, W.-C (2000) Data Provenance: Some Basic Issues. Springer Foundations of Software Technology and Theoretical Computer Science: pp. 87–93.

48. Amstutz P. CM, Tijanić N (editors), Chapman B., Chilton J., Heuer M., Kartashov A., Leehr D., Ménager H., Nedeljkovich M., Scales M., Soiland-Reyes S., Stojanovic L. (2016) Common Workflow Language, v1.0. Specification, Common Workflow Language working

49. Hettne KM, Dharuri H, Zhao J, Wolstencroft K, Belhajjame K, et al. (2014) Structuring research methods and data with the research object model: genomics workflows as a case study. J Biomed Semantics 5: 41.