bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

Gil Alterovits[2,3,4], Dennis Dean[9], Carole Goble[11], Michael R. Crusoe[16], Stian Soiland-Reyes[11], Amanda Bell[5,6], Anais Hayes[5,6], Charles Hadley King[5,6], Dan Taylor[20], Elaine Johanson[1], Elaine E. Thompson[1], Eric Donaldson[1], Hiroki Morizono[7, 22], Hsinyi S. Tsang[13, 14], Jeremy Goecks[10], Jianchao Yao[19], Jonas S. Almeida[8], Konstantinos Krampis[23, 24], Lydia Guo[17], Mark Walderhaug[1], Paul Walsh[15], Robel Kahsay[5,6], Srikanth Gottipati[21], Toby Bloom[12], Yuching Lai[18], Vahan Simonyan[1*], Raja Mazumder[5,6*]

1 US Food and Drug Administration, Silver Spring MD 20993, United States of America
2 Harvard/MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA
3 Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115, USA
4 Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Boston, MA 02139, USA,
5 Department of Biochemistry & Molecular Medicine, The George Washington University Medical Center, Washington, DC 20037, USA
6 McCormick Genomic and Proteomic Center, The George Washington University, Washington, DC 20037, USA
7 Department of Genomics and Precision Medicine, The George Washington University, Washington DC 20037, USA
8 Stony Brook University, School of Medicine and College of Engineering and Applied Sciences, Stony Brook, NY 11794, USA
9 Seven Bridges, Cambridge MA, 02142, USA
10 Computational Biology Program, Oregon Health & Science University, Portland OR, 97239, USA
11 School of Computer Science, The University of Manchester, Manchester, M13 9PL, UK
12 New York Genome Center, New York, NY 10013, 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 Nsilico Life Science, Nova Center, Belfield Innovation Park, University College Dublin, Dublin 4, Ireland
16 Common Workflow Language Project, Vlinius, Lithuania
17 Wellesley College, Wellesley, MA 02481, USA
18 DDL Diagnostic Laboratory, 2288 ER, Rijswijk, Netherlands
19 MRL IT, Merck & Co., Inc., Boston, MA, USA
20 Internet 2, 1150 18th St. NW, Washington, DC 20036, USA
21 Think Team, Otsuka Data Sciences, Ostuka Pharmaceutical Development and Commercialization, Inc. (OPDC), USA
22 Center for Genetic Medicine Research, Children's National Medical Center, Washington DC 20010, USA
23 Department of Biological Sciences, Hunter College of The City University of New York, USA
24 Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY 10021, USA

*Corresponding authors

Abstract. Precision medicine can be empowered by a personalized approach to patient care based on the patient's unique genomic sequence. To be used in precision medicine, genomic findings must be robust, reproducible, and experimental data capture should adhere to FAIR Data Guiding Principles. Moreover, precision medicine requires standardization that extends beyond wet lab procedures to computational methods.

Rapidly developing standardization technologies improves communication of genomic sequencing by introducing concepts such as error domain, usability domain, validation kit, and provenance information. These advancements allow data provenance to be standardized and ensure interoperability. Thus, a resulting bioinformatics computation instance that includes these advancements can be easily communicated, repeated and compared by scientists, regulators, clinicians and others, allowing a greater range of practical applications.

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. Through standardized bundling of HTS studies under an umbrella, regulatory agencies (FDA), academic researchers, and clinicians can expand collaboration to drive innovation in precision medicine with the potential for decreasing the time and cost associated with NGS workflow exchange, including FDA regulatory review submissions.

Keywords: BioCompute Objects, high-throughput sequencing, HTS. NGS, regulatory review, CWL, FHIR, GAG4H, HL7, and research objects

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

Genomics is a key tool for enabling precision medicine [1]. Thus, evidence based and 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). The ability to share sequence data and derived features for research provides insight into disease genetics, as demonstrated by large-scale projects like the Human Genome Project[2], The Cancer Genome Atlas (TCGA)[3], the 1000Genomes project[4, 5] and the International HapMap Project[6]. 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)[7] and ClinVar[8] as well as LD Hub, a centralized database of GWAS results for diseases/traits[9].

Modern methods to sequence and analyze large genomes have plummeted from high-throughput sequencing (HTS) costs. The cost for HTS has fallen from $20 per base in 1990 to less than $.01 per base in 2011, creating a mass accumulation of data[10]. Lower costs of HTS data generation increase the availability of data, expediting more and more types of analyses. Without a universal standard for communication, we quickly encounter the "Tower of Babel" problem of diversified languages and mass miscommunication. As such, a common HTS computational workflow standard for genomic sequencing bridges gaps and can be used in basic and clinical research diagnosis and prognosis, and in the development of companion diagnostics for novel therapeutics[11]. Furthermore, benchtop genome sequencers such as the small factor Illumina MiSeq or MiniSeq, are revolutionizing genomics by enabling access to HTS technology for smaller, independent laboratories in basic genetics and molecular biology research. This has resulted in generation of additional sequencing data by a large community of researchers, which are the fringe of the large-scale sequencing hubs, but still face similar bioinformatics bottlenecks in regards to analysis of their data.

Research in HTS-based methodologies has developed from sequence-based projects to more complex studies that examine genomes for genetic markers of diseases, vaccines, bacterial/viral strain identification, food contamination, etc.[12]. There is also a drive towards a multi-omics approach, where the integration of HTS data with other omics data sets can provide deeper insights into functional and /or systems biology of diseases and more accurate 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[13]. These initiatives allow different data sources to interact, advancing genomic analyses. However, standards for bioinformatic workflows should be established so information derived from data analyses, downstream of the original data producer, can be reproduced and validated for specific precision medicine use cases[14].

The U.S. Food and Drug Administration (FDA) have been working to standardize their computational and review processes for HTS and next generation sequencing (NGS) data. Regulation spurs innovation, improves regulatory decision-making, and provides safe and effective treatments[15]. Before gaining approval for clinical use, the FDA must clearly understand the bioinformatics analysis steps and computations, in case the analysis needs to be reproduced for a regulatory decision [16]. Therefore, the experimental results and the computational steps need to be understandable and reproducible to facilitate robust scrutiny and validation.

The National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH) has supported the creation of 64 CTSA Program Hubs. These hubs are designed to speed the transition of discoveries from bench to bedside and are encouraged to share tools and best practices amongst each other using mechanisms such as the Trial Innovation Network[17] and the SMART IRB platform[18]. The increased utilization of NGS in clinical trials both to stratify study populations and to identify the genomic bases of diseases means clearly verifiable and reproducible workflows are essential. The need for standardization becomes especially critical for integrating genomic and clinical information in patients with rare diseases. Another pressing need for establishing reproducibility is in performing meta-analyses. We now can create large scale data lakes of genomic information, and reprocessing raw data obtained from multiple contributors using identical workflows with the same toolsets ensures higher quality information.

Robustness and reproducibility depend not only on the wet lab protocols, but also on the computational workflows, pipelines, versions, environments, and parameters used in the process[19]. Rapid increases in availability of -*omic* data[20] and analysis of other biomedical data has created a bottleneck in downstream analysis, technological

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

advances, and critical community data communication[21]. To facilitate comprehension and comparison, standard reporting of detailed, traceable computational results needs to be implemented. These provenance pieces would be ideal for use across disciplines in publications, clinical trials, potential FDA submissions, and post-marketing surveillance. Capturing data provenance is crucial to improve methods of combining various sources and volumes of data efficiently and accurately[22]. Consistent reporting methods allow computational results to be shared more widely, save time and money previously used to reproduce results, and ensure validated methods include sufficient information for precision medicine applications[20].

Current standards already exist to capture genomic sequencing information and provenance: Fast Healthcare Interoperability Resources (FHIR)[23] and organizations such as Global Alliance for Genomics and Health[24] (GA4GH) communicate genomic information; these cater towards specific community domains. The Common Workflow Language (CWL)[25] and Research Objects (RO)[26] capture reproducible workflows in domain agnostic manner. Both Research Objects and FHIR use the W3C standard PROV to represent and interchange provenance information generated in different systems and under different contexts (https://www.w3.org/TR/prov-o/). While each of these models contribute to data sharing and generate robust and reproducible data, each holds a piece of the necessary description, requiring a universal framework.

BioCompute Object (BCO) unites these standards to provide framework of provenance reporting for genomic sequencing data analysis in the context of FDA submissions and regulatory review[3]. BCOs provide a new harmonizing approach designed to satisfy FAIR Data Principles[23] in the regulatory and research needs for evaluation, validation, and verification of bioinformatics pipelines[6, 11, 14].

In this paper, we will focus on how the FHIR, GA4GH, CWL, and RO standards can be leveraged and harmonized by examining the BCO framework. Once established, the BCO framework can be utilized for other types of FDA submissions, like large clinical trials, where data provenance in analysis datasets can be difficult to communicate. Utilizing the BCO framework will also ensure that when genomic pipelines are verified for accuracy, their provenance and appropriate uses will be easily known. As regulation for standards develops, reproducible data and interoperability become feasible for clinicians and researchers alike. Figure 1 provides a schematic of the BCO framework.

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

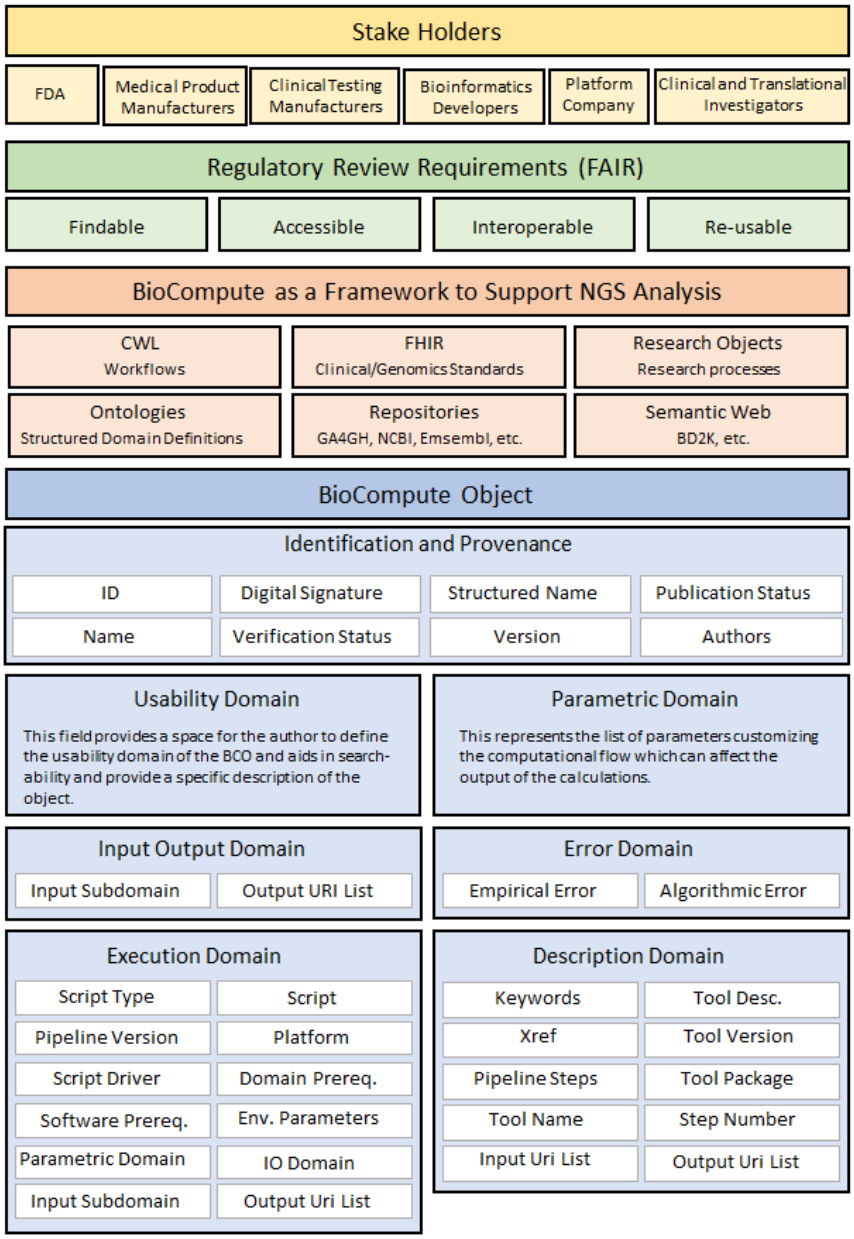

Fig 1. BioCompute Object as a Framework for advancing regulatory science by incorporating existing standards

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Background

The need for Reproducibility, Repeatability, and Record Keeping in Biomedical Experiments is necessary to facilitate interoperability among clinicians and researchers. To ensure validity and accuracy in experimental results, replication of data is necessary. With reproducible data processing, we can avoid unnecessary repetition of work and expedite regulatory approval like with FDA submissions[12] . The issue of reproducibility is highly relevant to current research. The Academy of Medical Sciences hosted a symposium to discuss reproducibility in pre-clinical biomedical research that identified multiple causes of reproducibility in research and identified a number of measures for improving reproducibility including creating greater openness and transparency, defining reporting guidelines, and better uses of standards and quality control measures [27]. Issues of reproducibility have resulted in enormous waste of research effort and have seriously frustrated progress in the life sciences, as highlighted several high profile articles[28] [29].

A distinction must be made between repeatability and reproducibility. The terms are similar in that both describe how closely repeated measurements of an entity can be expected to agree; both are crucial in designing a clinical trial, assessing the quality of laboratory procedures, and analyzing quantitative outcomes[30]. Repeatability and reproducibility differ in their testing conditions. While reproducibility uses different instruments to test a hypothesis, repeatability uses the same instruments for the technical replicates[30].

Reproducibility is the typical standard by which scientific claims are challenged[31]. With reproducibility, scientists can take repeated measurements of a physical quantity under realistic, variable conditions. For this reason, reproducibility is just as applicable to experiments or clinical trials as evaluations in analytical technologies and methods[32]. In measurement system analysis, repeatability is the number of standard deviations between measurements taken under the same conditions.

Some researchers distinguish categories of reproducibility, proposing two types of reproducibility: method reproducibility and result reproducibility[33]. 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 adhering to the methods closely. Method reproducibility depends significantly on the completeness of the procedures provided by the researchers. CWL creates a similar starting point for computational methods - if scientists gather data in the same computational language, then a large part of provenance model is standardized. As such, CWL and provenance tracking allow researchers to more easily 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 context of the experiment[34].

Researchers replicate results to eliminate spurious claims and enforce a disciplined approach, but replicating wet-bench work is often expensive and time-consuming. Complete repeatability is more feasible and superior in computational analyses [35]. For relatively lightweight workflows, such as the segmentation of moderately sized images in Pathology [36], a solution that is both repeatable and reproducible may be possible and open to collaborative identification of a computational object. While analytical repeatability is more straightforward, it still faces its own complications[37].

Record-keeping expectations for wet-lab and analytical research differ. Within physical, chemical, and biological experiments, environmental and procedural variability have been well accounted for[38]. But within computational work, parallel documentation accounting for the variability of parameters, versions, arguments, conditions, and protocols of computational biology algorithms and application usage has been documented much less rigorously. This lack of documentation is the largest hurdle to effective testing of a computational method[14]. Repeatability has proven to be an elusive standard for bioinformatics because of the lengthy and unstandardized nature of the studies conducted. For example, a study's processes could include sequencing biological samples, transferring extra-large data files to and from an archival server, and then repeating the computational workflows. Though a seemingly straightforward process, immense challenges to the repeatability of the experiment exist, including non-standardized file types and sizes, missing data provenance, incomplete computational workflows, outdated software versions, and missing/different parameters used. These factors culminate in an analysis that is nearly impossible to repeat[39].

These repeatability difficulties are manifested in practical applications like FDA submission approval. When FDA reviewers assess the validity of an experiment's results, the overall acceptance depends on the repeatability of the

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

data generation and subsequent bioinformatics analyses. For example, some FDA review divisions request that drug sponsors submit all NGS data and a detailed description of the experimental protocol. This includes the bioinformatics analysis pipelines used, so that the sequence information can be evaluated independently. The independent evaluation is to ensure it supports the claims made in the label of newly approved drugs. Currently, the repeatability of a NGS data analyses has been difficult due to the challenges described above. Internal analysis pipelines have therefore been employed in an attempt to reproduce the data and make comparisons to the results provided by the sponsor. By documenting pipeline analysis steps, tool information and parameters, BCOs enable reproducible and repeatable pipeline execution.

Universally reproducible data is an outcome to aim for, but current issues remain challenges. Without repeatability and analytics standards for NGS/HTS studies, regulatory agencies, academic researchers, pharmaceutical companies, and the FDA cannot work together to validate results and drive the emergence of new fields[39]. Without industry-wide standards to record computational workflow and the subsequent analysis, many studies are not repeatable and not usable in clinical or practical situations. To face these issues, workflow management systems and bioinformatics platforms have been developed to track and record computational workflows, pipelines, versions, and parameters used. However, these efforts remain haphazard and require BCOs to harmonize them. By tracking provenance of data and accurately documenting the trail of processes, BCOs enable reproducible data and increased information sharing.

Provenance of Data

To enable reproducible data, the origin information of the data and its trail must be captured in a standardized format. Data provenance refers to data's derivation history starting from the original sources, namely its *lineage*. Lineage graphs include the source of a piece of data into a database, data movement between databases or computational processes, or its generation from a computational process. The complement of data lineage is a *process audit* which provides a historical trail documenting the study, providing snapshots of intermediary states, values of configurations and parameters and of traceability of stepwise analytical processing[40]. Such audit trails should allow an independent reviewer to audit a computational investigation. Both gather provenance information that is crucial to ensure accuracy and validity of experimental results[41]. However, gathering such material is a challenge as modern web developments makes data transformation and copying easier, and computational workflows become capable of producing reams of fine-grained but not particularly useful trace records. For example, the molecular biology field supports 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[42]. Computational workflows can accurately collect lineage and process records, but obtaining appropriate granularity of the record keeping and 'black-box' steps in which inputs and outputs are not transparently connected or key processes are hidden remains a challenge[43, 44].

Issues with tracking data origins and the transparent and accurate record of executed process have far reaching effects in scientific work. Experiments rely on the confidence of the data's and the process's accuracy and validity, especially after undergoing complex, multistep processes of aggregation, modeling, and analysis[45]. Computational investigations require interactions with adjacent disciplines and disparate fields to effectively discover, access, integrate, curate, and analyze a large range and volume of relevant information. These hurdles require a solution beyond Open Data to establish Open Science in the community where provenance is preserved and shared to provide better transparency and reproducibility[46]. Because these complex analyses heavily rely 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 [47, 48] e.g. culminating in the W3C PROV[47], used by FHIR and ROs, based on the idea of generating an entity target via an agent's activity.

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

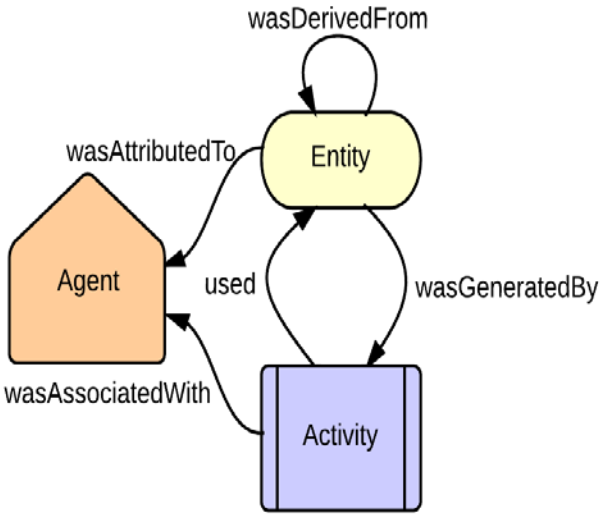

Fig 2. W3C Provenance Specification used in FHIR and ROs. Adapted from http://www.w3.org/TR/prov-primer/

Workflow management systems have developed to capture analysis processes and bioinformatics platforms to capture analysis details, gathering provenance information easily for users, extending PROV where necessary. However, greater standardization of these systems is needed and comes in the form of BCOs, which ensure provenance is recreated for regulatory approval and that standards such as PROV are appropriately adopted.

Workflow Management Systems

Scientific workflows have emerged as a model to represent and manage complex distributed scientific computations [22]. 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 to the analysis' steps and data transformations, workflows also capture the mechanisms to carry out the steps in a distributed environment, as in the use of specific execution and storage resources in this computing environment. A significant amount of work has been done to share workflows and experiments [49, 50]. 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 among collaborators[51].

Workflow management systems act to execute and monitor scientific workflows, coordinating the sequential components [52]. Developments in workflow management systems have led to the proposition of using workflow-centric research objects with executable components[13, 34]. 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 [15, 53]. Extensive reviews of current workflow systems for bioinformatics are linked [16, 53-55]. Ongoing systems participate in the current trend of moving from graphical system back to script-like workflows. These systems are now executed on cloud infrastructure, 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 mostly not in the PROV standard. Therefore, BCOs rely on existing regulatory standards like CWL to manage pipeline details; and ROs and FHIR to unify and draw data from workflows to enhance interoperability.

Bioinformatics platforms

The dramatic increase in NGS technology use has resulted in the rapid increase in scalability needs to store, access, and compute reads and other NGS/biomedical data[56]. These technical and physical requirements have led to a field-wide call for integrated storage and computational node usage methods. Such integration will minimize data transfer costs and remove the bottlenecks found in both downstream analyses and community communication of

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

computational analyses results[57] to build upon existing knowledge. For bioinformatics platforms, communication requirements include (a) recording all analysis details such as parameters and input datasets; and (b) sharing analysis details so others can understand and reproduce analyses.

To reduce the unprocessed data buildup, several high-throughput [20, 22] cloud-based infrastructures have developed including HIVE (the High-performance Integrated Virtual Environment)[57], and Galaxy[58], along with commercial platforms from companies like DNAnexus, and Seven Bridges Genomics, among others. High throughput computing (HTC) environments deliver large amounts of processing capacity over long periods of time, an ideal environment for long-term computation projects, as with genomic research[59]. Current platforms utilize distributed cloud computing environments to support extra-large dataset storage and computation, and host tools and workflows for germline and somatic variant calling, RNA-seq, microbiome characterization, and many more common analyses. These cloud-based infrastructures and tools reduce data silos, converting the data to reproducible formats to 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 data sets available on TCGA[3].

Overall, the genomic community has come to acknowledge the necessity of data sharing and communication to facilitate reproducibility and standardization[12]. Data sharing is crucial in situations ranging from long term clinical treatments to the ability to respond to public health emergencies[60]. As the infrastructure, community accessible resources, and data sharing industry policies develop, the need for voluntary, industry-wide standardization across a wide range of stakeholders becomes undeniable to ensure that published results are reproducible and robust. Extending bioinformatics platforms to include data provenance, standard workflow computation and encoding results with available standards through implementation of the BCO would greatly support the exchange of genomic data analysis methods for regulatory review.

Discussion

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, a clinical trial or any submission supported with NGS results can require considerable time and expertise to review. Submission of a BioCompute Object (BCO) would ensure that data provenance is unambiguous and that the bioinformatics workflow is fully documented [41, 42, 46].

To truly understand and compare computational tests, a standard method (like BCO) requires tools to capture progress and communicate the workflow and input/output data. As the regulatory field progresses, the following methods have been developed and are continually refined to capture workflows and exchange data electronically[22].

Biocompute Objects (BCOs) and Their Harmonizing Efforts

Biocompute Objects (BCOs) were conceptualized to alleviate the disparate nature of HTS computational analysis. The primary objective of BCOs is to (a) harmonize HTS 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[35]. Each BCO is comprised of 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 conceptual schema for BCO creation is built on top of two layers: the data definition framework and the BCO framework [38, 39, 61].

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

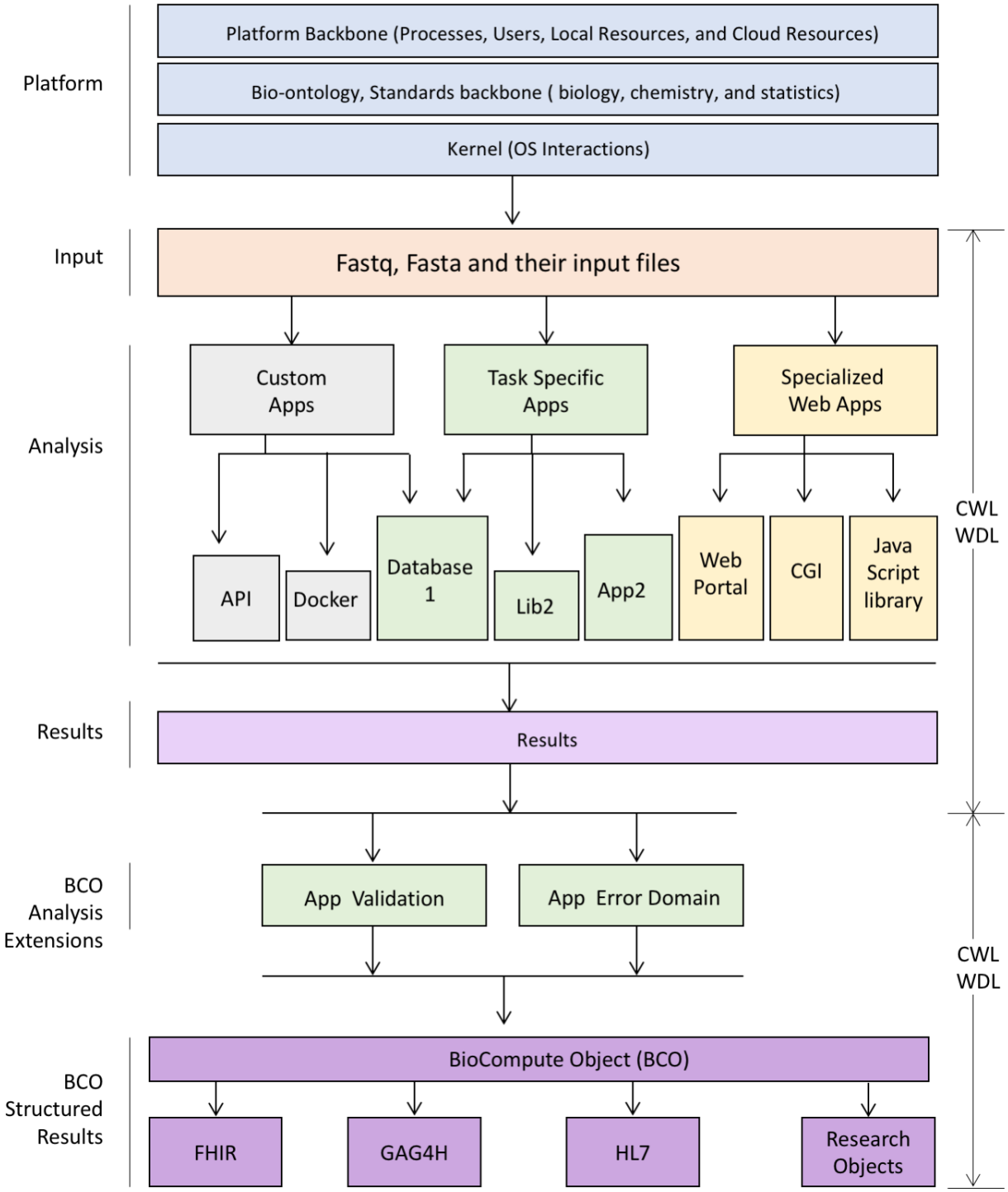

Fig 3. Generic HTS Platform schematic with proposed BioCompute Object extensions.

The data definition framework contains primitive data type definitions and categorizes them as an atomic type. In this case an atomic type is that which cannot be deconstructed any further without losing meaning or some other important information, like an integer or character. Complex types are composed of multiple atomic types or even

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

multiple complex types, like a character string. Using these principles, one can construct a datum that has the ability to represent any level of complexity needed, with the only constraint being the amount of available storage memory or computing power.

When defining a field in a data type, one can place any number of constraints on the data that the field will accept as valid. If one were constructing a data type field to hold DNA sequencing information, one could restrain the type of characters that field would accept. This further refinement ensures that only the characters used to represent nucleic acids would be accepted as input in this field (A, T, C, and G).

The second framework layer defines a derived data type called the "primitive biocompute type." Extending the same principles that allowed one to construct a string representing a DNA sequence from the primitive character type, it is possible to construct a data type definition with the absolute minimum fields necessary to create a BCO. By taking the primitive BioCompute object type and adding parametric and metadata fields unique to a particular instance, one can get the final, unique BCO for the specified workflow and analysis.

The declarative nature of BCOs suggests an implementation with minimal procedural barriers. Though not a requirement, the use of schemaless representation in JSON format that does not impede the identification of a validating schema accords with the purpose of Compute Objects such as BCOs and FHIR I/O. For an example of editable BCO objects[51], see https://mathbiol.github.io/bco. Accessing experimental data and its origin is challenging; thus, aligning frameworks that encourage interoperability such as Data Tag Suite (DATS) help attain data standards that are easily verifiable, discoverable, reusable and interoperable[62]. DATS is a mechanism that enables the data to be easily searchable, findable, and reusable. The BCO takes a snapshot of the whole experiment computational procedure where the input data is provided and described in detail along with all the default and experimental procedures used in the dataset. The output domain of the BCO includes the results from the experiment in the dataset so that any other user can run the exact experiment and produce the same results. The BCO captures curated ontologies which are in reviewed and highly maintained databases to ensure that they are easily accessible and searchable.

The BCO can serve as an umbrella of standards allowing for standards such as Common Workflow Language (CWL), Fast Healthcare Interoperability Resources (FHIR), Global Alliance for Genomics and Health (GA4GH), and Research Objects (RO) to be embedded within BioCompute Object fields. Enabling BCOs to incorporate existing standards provides a universal framework for including existing advances in workflow and data specifications that greatly increase the specificity for which to describe a workflow and the related provenance. Moreover, the umbrella approach also supports a minimal effort form based BCO that can be quickly implemented allowing for a rapid initial implementation that can evolve overtime to capture the greater specificity made available by incorporating existing standards.

The Common Workflow Language (CWL)

Common Workflow Language (CWL)[63] is an open community-led standard to describe workflow and tools for data-intensive sciences (including Bioinformatic and Medical Imaging analyses) with a strong focus on reproducibility, reusability, scalability and portability. CWL files can be executed by multiple workflow engine implementations, including Toil, Arvados, and RabixBunny[64]. These implementations again support execution locally, on clusters, and on multiple cloud and HPC environments.

In an effort to standardize, CWL has focused on the current ability of most workflow systems: *Execute command line tools* and *coordination of their inputs and outputs* in a top-to-bottom pipeline. At the heart of CWL workflows are WL tool descriptions. A command line, often with an accompanying Docker container, is described with parameters; and linking to and from registries like ELIXIR's (European Life-sciences Infrastructure for Biological Information) bio.tools [65]. These are then wired together in another YAML file to form a workflow template, which can be executed repeatedly on any supported platform by specifying input files and workflow parameters.

CWL allows scientists to express their data and workflows in a universal computational language, generating greater method reproducibility for the genomic community. A community-specific computing language builds standardization from the data producer, avoiding the "Tower of Babel" issue of varied languages causing miscommunication. CWL lays the foundation for expression of BCOs, inherently embedding reproducibility in the BCO specification.

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Fast Healthcare Interoperability Resources (FHIR)
FHIR is an all-encompassing standard for communicating clinical and health information. As such, it includes genomic components known as FHIR Genomics API/specification integrated in its core. These genomics components evolved from the SMART on FHIR Genomics standard[66] and integrated work of the HL7 Clinical Genomics Workgroup, and the standard is based on the requirements of Meaningful Use 3.0. FHIR is an emerging standard for Electronic Medical Records (EMRs) and clinical apps being adopted by numerous vendors in the healthcare space. Projects based on FHIR Genomics enable lab vendors to share clinical genomic information for precision medicine and EMR-based patient information for research studies, such as the NIH's All of Us program. Projects based on FHIR enable both the data and ecosystem to exist for communication of clinical and genomic information on individual patients.

Capturing genomic provenance information via FHIR enables clinical trials, research, and clinical interpretations to be traceable back to the original methods, workflows, and parameters used. This, in turn, facilitates robust and reproducible clinical interpretations of genomics and comparisons to be made across patients in which similar methods were used. FHIR utilizes the PROV standard introduced earlier to capture provenance information. Practically applied, a clinical genomic sequence entity target can be generated via a particular workflow instance activity through a specific laboratory agent. As part of the FHIR Release 3 API/specification, provenance examples are constructed that enable the capture of workflows via CWL and workflow instance for potential FDA submissions via BCO. FHIR equips clinicians, researchers, and regulators to be able to trace, reproduce, and reinterpret/compare genomic data[67]. By communicating clinical information, FHIR lays the groundwork for collaboration in BCO implementation, permitting easy sharing of data.

Global Alliance for Genomic Health (GA4GH)
The Global Alliance for Genomic Health (GA4GH) is a cooperative framework established as a resource for genomic research and phenotype sharing [68]. GA4GH was created as a common framework to enable responsible, voluntary, and secure sharing of data to advance precision care[69]. It has faced challenges in data aggregation procedures, but has demonstrated the potential of a synergistic data sharing culture. To execute data sharing goals, GA4GH schemas, which define how to access genomic data, and APIs, which implement these schemas, have been created. These schemas facilitate DNA sequence data exchange and use common, user-friendly web protocols to overcome incompatible infrastructures [68]. An application of GA4GH is the BRCA Exchange (http://brcaexchange.org/), which provides a searchable resource that combines breast cancer-contributing germline variants from eight different institutions. Overall, GA4GH is not intended to enforce data standards, but rather provides recommendations to influence and persuade the advantages of a collaborative data culture[70]. GA4GH enables researchers to communicate their data to clinicians and the FDA. Together with FHIR, GA4GH allows BCOs to be utilized in clinical and basic research, and as BCOs are integrated with these specifications, data communication and provenance information become interlinked.

Research Objects (ROs)
Research Objects [http://ResearchObject.org/] is a new publication model that improves reproducibility of scientific data by capturing provenance, quality, credit, attribution, and methods[26, 30]. A Research Object (RO) is an aggregation mechanism that bundles the method of a computational analysis (e.g. expressed as scripts and workflows) and all associated materials, metadata, and annotations using existing Linked Data standards[71]. ROs consist of a container of files with a manifest to provide meaningful information about what those files are, what they mean, how they relate and provide provenance and versioning information [11]. The containers vary, such as *Docker, BagIt[72]* or the Zip Archive *Research Object Bundle* [https://w3id.org/bundle/]. Resource content can be embedded or referenced externally using URIs, which may require further authentication and allows for greater regulation. ROs collect the general data and workflow provenance necessary for reproducibility, acting as a lab notebook for computational processes.

ROs have been applied to improve reproducibility of workflows [73] and to describe large datasets [31, 72]. By its aggregating nature, ROs go beyond the experimental description to bring together the wider digital context of scientific processes and their conduct, including input/output data, methods, software, actors, analysis, dissemination, sharing, reuse, and the links/relationships between these gathered resources[74].

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The *wf4ever* project[75], which primarily developed the RO model[34], specified a workflow description vocabulary (*wfdesc*) [https://w3id.org/ro/2016-01-28/wfdesc] that defines resources associated with a workflow specification within a Research Object Framework. The workflow description vocabulary defines three main terms: workflow as a process node and data link, process as a software tool that executes specific actions, and data link as a tool used to encode dependencies between computational nodes. This vocabulary is the basis for the CWL specification[63]: CWL provides the description of the workflow and its means of execution whereas the RO provides the description relating the workflow to its provenance, purpose and so forth. The PROV resource (Figure 2) ontology is also the basis of the RO workflow provenance model *wfprov* [https://w3id.org/ro/2016-01-28/wfprov], linking the various specifications (CWL, FHIR, ROs) under a similar basis that provides interoperability leveraged by BCOs.

Journal/Peer Review Perspective

The genomic community has come to acknowledge the necessity of data sharing and communication to facilitate reproducibility, standardization and provenance, reshaping the way research is conducted, ensuring openness and maximum benefit by the scientific community who ultimately is the consumer of the products of a research publication[12].

This issue is clearly exemplified by the lack of interoperability between the web service interfaces of major bioinformatics centers, including the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) in the UK, DNA Data Bank of Japan (DDBJ)/Kyoto Encyclopedia of Genes, and Genomes (KEGG)/Protein Data Bank Japan (PDBj) in Japan. As the centers' web service models are all based on open standards, their databases and computational resources are expected to be interoperable[76]. Despite the large amount of data in these services, these centers use their own data type definitions, making it harder for end users and developers to utilize these services to create biological analysis workflows[77].

While lack of interoperability is not uncommon in computational biology, significant efforts have been made to increase interoperability between web services, standardize exchangeable data types, and adopt compatible interfaces for each service[78]. Several projects and workshops have already begun progress to bridge the gap: the BioMoby project defined ontologies for data types and methods used in its services, and it provides a centralized repository for its service discovery[79]; Open Bio* libraries have been developed for the major computing languages i.e. Perl, Python, Ruby, and Java) to maximize bioinformatics web services and to create collaborative compatible data models for common biological objects[80]; the EDAM ontology of bioinformatics operations, types of data and identifiers, topics and formats used by CWL and workflow ROs[81] the DBCLS BioHackathon improves web service interoperability and collaboration between major database centers[78]; and the HTS-CSRS Workshop hosted by GWU and the FDA is a cross-disciplinary endeavor that emphasizes standardization of data storage and collection, communication of this genetic data, and the necessity of reproducibility of these analyses to ensure their potential clinical applications[32]. These are just a few examples of the efforts that combine technologies, ontologies, and standards to enhance computational analysis information. The FAIRsharing.org portal (formally biosharing.org) for metadata standards in the biosciences has a comprehensive curated catalogue [82]. The positive response to improving interoperability indicates the community's need for such standardization[22]).

Conclusion

Robust and reproducible data analysis is key to successful precision medicine and genome initiatives. Researchers, clinicians, administrators, and patients are all tied by the electronic information in EHRs and databases. Current systems rely on data stored with incomplete provenance records and varying computing languages, creating a cumbersome and inefficient healthcare environment.

The initiatives discussed 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 data silos of unusable data, making standardized regulation of reproducibility more dire. To open the bottleneck of downstream analysis, the provenance (or origin) of data plus analysis details (e.g., parameters, workflow versions), must be tracked to ensure accuracy and validity. Developing high-throughput cloud-based infrastructures like DNAnexus, Galaxy, HIVE, and Seven Bridges Genomics can capture data provenance and store the analyses in infrastructures that allow easy user interaction.

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 clinical and research settings in both the conceptual experimental methods and the underlying computational workflows. There are several use cases of such work including submission for FDA diagnostics evaluations, as is being done with the BCO effort. Such standards will also enable robust and reproducible science to facilitate open science between collaborators. At this juncture, it is imperative to lead the development and improvement of these standards to satisfy needs of downstream consumers of genomic information to validate and reproduce key workflows and analyses.

The need to communicate HTS/NGS computations and analyses reproducibly has led to increased collaboration among disparate players in industry through conferences/workshops that increase exposure to standardization, tracking, and reproducibility methods [63, 71]. 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) can also increase the scalability and reproducibility of data analysis. BCOs act a harmonizing umbrella to facilitate data submitted to regulatory agencies, increasing interoperability in the genomic field. BCOs are easily generated by bioinformatics platforms that automatically pull underlying data and analysis provenance into their infrastructures. 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 makes, robust and reproducible analyses and results an attainable standard for the scientific community.

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 the "Enabling Precision Medicine via standard communication of NGS provenance, analysis, and results" paper. 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 by award number UL1TR001876 from the NIH National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences or the National Institutes of Health.

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

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Ginsburg, G.S. and H.F. Willard, *Genomic and personalized medicine: foundations and applications*. Transl Res, 2009. **154**(6): p. 277-87.
2. Haq, M.M., *Medical genetics and the Human Genome Project: a historical review*. Tex Med, 1993. **89**(3): p. 68-73.
3. Tomczak, K., P. Czerwinska, and M. Wiznerowicz, *The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge*. Contemp Oncol (Pozn), 2015. **19**(1A): p. A68-77.
4. Abecasis, G.R., et al., *A map of human genome variation from population-scale sequencing*. Nature, 2010. **467**(7319): p. 1061-73.
5. Auton, A., et al., *A global reference for human genetic variation*. Nature, 2015. **526**(7571): p. 68-74.
6. Manolio, T.A., L.D. Brooks, and F.S. Collins, *A HapMap harvest of insights into the genetics of common disease*. J Clin Invest, 2008. **118**(5): p. 1590-605.
7. Mailman, M.D., et al., *The NCBI dbGaP database of genotypes and phenotypes*. Nat Genet, 2007. **39**(10): p. 1181-6.
8. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype*. Nucleic Acids Res, 2014. **42**(Database issue): p. D980-5.
9. Zheng, J., et al., *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, 2017. **33**(2): p. 272-279.
10. Sawyer, E., *High Throughput Sequencing and Cost Trends*. Nature Education, 2017.
11. Boyd, S.D., *Diagnostic applications of high-throughput DNA sequencing*. Annu Rev Pathol, 2013. **8**: p. 381-410.
12. Kaye, J., et al., *Data sharing in genomics--re-shaping scientific practice*. Nat Rev Genet, 2009. **10**(5): p. 331-5.
13. in *Evolution of Translational Omics: Lessons Learned and the Path Forward*, C.M. Micheel, S.J. Nass, and G.S. Omenn, Editors. 2012: Washington (DC).
14. Simonyan, V., J. Goecks, and R. Mazumder, *Biocompute Objects-A Step towards Evaluation and Validation of Biomedical Scientific Computations*. PDA J Pharm Sci Technol, 2017. **71**(2): p. 136-146.
15. Woodcock, J. and R. Woosley, *The FDA critical path initiative and its influence on new drug development*. Annu Rev Med, 2008. **59**: p. 1-12.
16. Xu, J., et al., *The FDA's Experience with Emerging Genomics Technologies-Past, Present, and Future*. AAPS J, 2016. **18**(4): p. 814-8.
17. NCATS, N. *Trial Innovation Network*. 2017 06/12/2017 [cited 2017 10/19/2017]; Available from: https://ncats.nih.gov/ctsa/about/network.
18. SMART-IRB. *Use SMART IRB to enable single IRB review*. 2017 [cited 2017 10/18/2017]; Available from: https://smartirb.org/.
19. Stodden, V., et al., *Enhancing reproducibility for computational methods*. Science, 2016. **354**(6317): p. 1240-1241.
20. Desai, N., et al., *From genomics to metagenomics*. Curr Opin Biotechnol, 2012. **23**(1): p. 72-6.
21. Scholz, M.B., C.C. Lo, and P.S. Chain, *Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis*. Curr Opin Biotechnol, 2012. **23**(1): p. 9-15.
22. Gil, Y., et al., *Examining the challenges of scientific workflows*. Computer, 2007. **40**(12): p. 24-+.

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

23. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data, 2016. **3**: p. 160018.
24. Lawler, M., et al., *All the World's a Stage: Facilitating Discovery Science and Improved Cancer Care through the Global Alliance for Genomics and Health*. Cancer Discov, 2015. **5**(11): p. 1133-6.
25. Amstutz P., C.M.R., Tijanić N (editors), Chapman B., Chilton J., Heuer M., Kartashov A., Leehr D., Ménager H., Nedeljkovich M., Scales M., Soiland-Reyes S., Stojanovic L. . *Common Workflow Language, v1.0. Specification, Common Workflow Language working* 2016.
26. Bechhofer, S., et al., *Why linked data is not enough for scientists*. Future Generation Computer Systems-the International Journal of Grid Computing and Escience, 2013. **29**(2): p. 599-611.
27. Anon, *Reproducibility and reliability of biomedical research*. The Academy of Medical Sciences.
28. Pusztai, L., C. Hatzis, and F. Andre, *Reproducibility of research and preclinical validation: problems and solutions*. Nat Rev Clin Oncol, 2013. **10**(12): p. 720-4.
29. Samuel Reich, E., *Cancer trial errors revealed*. Nature, 2011. **469**(7329): p. 139-40.
30. Connett, J., *Repeatability and Reproducibility, with Applications to Design of Clinical Trials*. Wiley Online Library, 2008.
31. Peng, R.D., *Reproducible research in computational science*. Science, 2011. **334**(6060): p. 1226-7.
32. Garijo, D., et al., *Quantifying reproducibility in computational biology: the case of the tuberculosis drugome*. PLoS One, 2013. **8**(11): p. e80278.
33. Goodman, S.N., D. Fanelli, and J.P. Ioannidis, *What does research reproducibility mean?* Sci Transl Med, 2016. **8**(341): p. 341ps12.
34. Belhajjame, K., *Using a suite of ontologies for preserving workflow-centric research objects*. Web Semantics: Science, Services and Agents on the World Wide Web, 2015(32(C)): p. pp.16–42.
35. Tabb, D.L., et al., *Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry*. J Proteome Res, 2010. **9**(2): p. 761-76.
36. Almeida, J.S., et al., *ImageJS: Personalized, participated, pervasive, and reproducible image bioinformatics in the web browser*. J Pathol Inform, 2012. **3**: p. 25.
37. Freire, J., Bonnet, P. & Shasha, D., *Computational Reproducibility: State-of-the-art, Challenges, and Database Research Opportunities*. SIGMOD, 2012. **Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data**: p. pp. 593–596.
38. Deyo, R.A., P. Diehr, and D.L. Patrick, *Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation*. Control Clin Trials, 1991. **12**(4 Suppl): p. 142S-158S.
39. Kjer, K.M., J.J. Gillespie, and K.A. Ober, *Opinions on multiple sequence alignment, and an empirical comparison of repeatability and accuracy between POY and structural alignment*. Syst Biol, 2007. **56**(1): p. 133-46.
40. Bose, R. and J. Frew, *Lineage retrieval for scientific data processing: A survey*. Acm Computing Surveys, 2005. **37**(1): p. 1-28.
41. Buneman, P., Khanna, S. & Tan, W.-C, *Data Provenance: Some Basic Issues*. Springer, 2000. **Foundations of Software Technology and Theoretical Computer Science**: p. pp. 87–93.
42. Buneman, P., Khanna, S. & Wang-Chiew, T, *Why and Where: A Characterization of Data Provenance. In Database Theory*. Springer, 2001. **Lecture Notes in Computer Science**(International Conference on Database Theory): p. pp. 87–93.

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

43. Freire, J., Bonnet, P. & Shasha, D. *Computational Reproducibility: State-of-the-art, Challenges, and Database Research Opportunities*. in *In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 2012. New York, NY, USA: SIGMOD.
44. Alper, P. *Enhancing and Abstracting Scientific Workflow Provenance for Data Publishing*. in *Joint EDBT/ICDT 2013 Workshop*. 2013.
45. Miles, S., *The Requirements of Using Provenance in e-Science Experiments*. Journal of Grid Computing, 2007. **5**(1): p. pp.1–25.
46. Reichman, O.J., M.B. Jones, and M.P. Schildhauer, *Challenges and Opportunities of Open Data in Ecology*. Science, 2011. **331**(6018): p. 703-705.
47. Moreau, L., et al., *The Open Provenance Model core specification (v1.1)*. Future Generation Computer Systems-the International Journal of Grid Computing and Escience, 2011. **27**(6): p. 743-756.
48. Ciccarese, P., et al., *PAV ontology: provenance, authoring and versioning*. J Biomed Semantics, 2013. **4**(1): p. 37.
49. Garijo, D., Y. Gil, and O. Corcho, *Abstract, link, publish, exploit: An end to end framework for workflow sharing*. Future Generation Computer Systems-the International Journal of Escience, 2017. **75**: p. 271-283.
50. Goble, C.A., et al., *myExperiment: a repository and social network for the sharing of bioinformatics workflows*. Nucleic Acids Res, 2010. **38**(Web Server issue): p. W677-82.
51. Deelman, E., et al., *Workflows and e-Science: An overview of workflow system features and capabilities*. Future Generation Computer Systems-the International Journal of Grid Computing-Theory Methods and Applications, 2009. **25**(5): p. 528-540.
52. Atkinson, M., et al., *Scientific workflows: Past, present and future*. Future Generation Computer Systems-the International Journal of Escience, 2017. **75**: p. 216-227.
53. Cohen-Boulakia, S., et al., *Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities*. Future Generation Computer Systems-the International Journal of Escience, 2017. **75**: p. 284-298.
54. Leipzig, J., *A review of bioinformatic pipeline frameworks*. Brief Bioinform, 2017. **18**(3): p. 530-536.
55. Spjuth, O., et al., *Experiences with workflows for automating data-intensive bioinformatics*. Biol Direct, 2015. **10**: p. 43.
56. Metzker, M.L., *Sequencing technologies - the next generation*. Nat Rev Genet, 2010. **11**(1): p. 31-46.
57. Simonyan, V. and R. Mazumder, *High-Performance Integrated Virtual Environment (HIVE) Tools and Applications for Big Data Analysis*. Genes (Basel), 2014. **5**(4): p. 957-81.
58. Afgan, E., et al., *Harnessing cloud computing with Galaxy Cloud*. Nat Biotechnol, 2011. **29**(11): p. 972-4.
59. Thain, D., T. Tannenbaum, and M. Livny, *Distributed computing in practice: the Condor experience*. Concurrency and Computation-Practice & Experience, 2005. **17**(2-4): p. 323-356.
60. Whitty, C.J., *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, 2017. **372**(1721).
61. Tong, W., et al., *ArrayTrack--supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research*. Environ Health Perspect, 2003. **111**(15): p. 1819-26.
62. Sansone, S.A., et al., *DATS, the data tag suite to enable discoverability of datasets*. Sci Data, 2017. **4**: p. 170059.
63. Amstutz, P., *Common Workflow Language, v1.0*. figshare, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

64. Kaushik, G., et al., *Rabix: An Open-Source Workflow Executor Supporting Recomputability and Interoperability of Workflow Descriptions*. Pac Symp Biocomput, 2016. **22**: p. 154-165.
65. Ison, J., et al., *Tools and data services registry: a community effort to document bioinformatics resources*. Nucleic Acids Res, 2016. **44**(D1): p. D38-47.
66. Alterovitz, G., et al., *SMART on FHIR Genomics: facilitating standardized clinico-genomic apps*. J Am Med Inform Assoc, 2015. **22**(6): p. 1173-8.
67. Michener, W.K., *Meta-information concepts for ecological data management*. Ecological Informatics, 2006. **1**(1): p. 3-7.
68. Lawler, M., et al., *All the World's a Stage: Facilitating Discovery Science and Improved Cancer Care through the Global Alliance for Genomics and Health*. Cancer Discovery, 2015. **5**(11): p. 1133-1136.
69. Hayden, E.C., *Geneticists push for global data-sharing*. Nature, 2013. **498**(7452): p. 16-17.
70. Siu, L.L., et al., *Facilitating a culture of responsible and effective sharing of cancer genome data*. Nature Medicine, 2016. **22**(5): p. 464-471.
71. Hettne, K.M., et al., *Structuring research methods and data with the research object model: genomics workflows as a case study*. J Biomed Semantics, 2014. **5**(1): p. 41.
72. Chard, K., et al., *I'll Take That to Go: Big Data Bags and Minimal Identifiers for Exchange of Large, Complex Datasets*. 2016 Ieee International Conference on Big Data (Big Data), 2016: p. 319-328.
73. Gonzalez-Beltran, A., et al., *From Peer-Reviewed to Peer-Reproduced in Scholarly Publishing: The Complementary Roles of Data Models and Workflows in Bioinformatics*. PLoS One, 2015. **10**(7).
74. De Roure, D., *Towards the preservation of scientific workflows*, in *International Conference on Preservation of Digital Objects* 2011: UK.
75. Page K, P.R., *From workflows to Research Objects: an architecture for preserving the semantics of science*, in *Procedings of the 2nd International Workshop on Linked Science* 2012: Boston, USA.
76. Stein, L., *Creating a bioinformatics nation*. Nature, 2002. **417**(6885): p. 119-20.
77. Navas-Delgado, I., et al., *Intelligent client for integrating bioinformatics services*. Bioinformatics, 2006. **22**(1): p. 106-11.
78. Katayama, T., et al., *The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. The DBCLS BioHackathon Consortium**. J Biomed Semantics, 2010. **1**(1): p. 8.
79. Wilkinson, M.D. and M. Links, *BioMOBY: an open source biological web services proposal*. Brief Bioinform, 2002. **3**(4): p. 331-41.
80. Stajich, J.E. and H. Lapp, *Open source tools and toolkits for bioinformatics: significance, and where are we?* Brief Bioinform, 2006. **7**(3): p. 287-96.
81. Ison, J., et al., *EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats*. Bioinformatics, 2013. **29**(10): p. 1325-32.
82. McQuilton, P., et al., *BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences*. Database (Oxford), 2016. **2016**.

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Appendix

Workflow Management Systems

bcbio-nextgen [https://github.com/chapmanb/bcbio-nextgen] has a domain-specific language for executing pipelines in HTS analysis, in particular variant calling like RNA-seq and small RNA analysis. Unlike other systems, bcbio-nextgen focuses on the parameters of the pipeline and a choice of algorithms, rather than the declaration of the steps and their underlying command lines. Bcbio handles installation of all third party tools and reference datasets required for its pipelines. Pipelines can be executed using multiple cores or parallel messaging on a cluster environment, which can facilitate high performance schedulers like LSF and SGE.

Snakemake [doi:10.1093/bioinformatics/bts480] is a declarative Python-like workflow language similar to a traditional *Makefile*. Snakemake files contain rules on how to create a particular file by executing a command or script and declaring which other files or file patterns the rule depend on, thus implicitly containing the rule execution order. The integration with Python simplifies "shim" operations between steps (e.g., handling different genomics file formats). The resulting workflow can be effectively executed on a local single-core machine, a multi-core server, or scaled to compute-clusters of different architectures.

Nextflow [doi:10.1038/nbt.3820] is a Python-like language for data-driven computational bioinformatics pipelines, with a strong focus on reproducibility and scalability. Nextflow uses Docker [https://www.docker.com/] to containerize and deploy the third-party tools the workflow relies on. A Nextflow workflow is declared by defining processes, which consume and produce messages on asynchronous channels. Channels are then wired together to form a workflow, which can be executed efficiently on a multitude of HPC and cloud platforms, including SGE, LSF, SLURM, Apache Ignite and Kubernetes.

Toil [doi:10.1038/nbt.3772] can run large-scale scientific workflows on cloud and HPC environments defined in either Common Workflow Language (CWL), Workflow Description Language (WDL), or Python Toil scripts. Toil jobs can be containerized using Docker and executed on multiple cloud environments (like AWS, Microsoft Azure, Google Cloud), in HPC environments using Grid Engine, or on distributed systems using Apache Mesos, with a strong emphasis on scalability and portability.

Bioinformatics Platforms

DNAnexus

Founded in 2009, DNAnexus (www.dnanexus.com) is a global, cloud-based platform for genomic data analysis and management. To meet increasing demands for efficient DNA data organization, DNAnexus arose as a tool for quick analysis of innumerable raw sequencing data, secure integration of genomic data with clinical infrastructures, and increased collaboration among scientists. The platform allows users to custom, port, and reproduce pipelines to the cloud-based infrastructure, making the data easily accessible. DNAnexus ensures clinically compliant data is secure and auditable. Additionally, DNAnexus facilitates collaboration among colleagues and upstream/downstream partners, easing data sharing.

Galaxy

Started in 2005, Galaxy (https://galaxyproject.org/) is an open-source, web-based platform that enables scientists without informatics expertise to perform computational analysis through the Web (Afgan et al. 2016). Existing analysis tools are integrated into Galaxy and are available through the consistent Web interface that can be deployed on any Unix system. Because the Galaxy software is highly customizable, the platform integrates with a wide variety of compute environments, making data processing accessible among users. Automated, multi-step analyses can be performed by combining tools into workflows (pipelines), and all analyses are reproducible (Goecks et al. 2010)(Blankenberg et al. 2010). By bridging the gap between tool developers and scientists, Galaxy helps both constituencies accelerate their research. The Public Galaxy Server (https://usegalaxy.org/) is an installation of the Galaxy software combined with many common analysis tools, workflows, and data sources. A free resource, the site provides substantial compute resources to analyze large datasets, transforming data to reproducible formats. The Galaxy Tool Shed (https://usegalaxy.org/toolshed) facilitates sharing of Galaxy tools as a central location where developers can upload their tool configurations, allowing greater collaboration for computational analyses. Galaxy formats data to be stored, imported, and exported for analyses and open workflows. Galaxy predates the implementation of community standards like GA4GH schemas, CWL, and BioCompute Objects, so the platform

bioRxiv preprint doi: https://doi.org/10.1101/191783; this version posted October 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

provides limited support for data standardization. Future developments should standardize Galaxy's data and methods to comply with current community standards.

Hive

The HIVE (Scholz et al. 2012; Simonyan et al. 2017; Metzker 2010; Simonyan & Mazumder 2014) platform is a cloud infrastructure that hosts a web-accessible interface that allows users to interact (deposit, share, retrieve, annotate, compute, visualize) with large volumes of NGS data. User interaction is conducted in a scalable fashion through the platform's connected distributed storage library and distributed computational resources. A novel aspect of HIVE compared to existing technologies is the seamless integration, hierarchical sharing, secure object traceability and auditing, presence of HIVE and existing external algorithms, biocuration, and FDA regulatory compliance (Logares et al. 2012; Whitty 2017). This platform allows users to regulate, reproduce, share and access data, and store computational workflows, complete with input/output data, parameters, versions, and tool usage (Simonyan & Mazumder 2014).

Seven Bridges Genomics (SBG)

Seven Bridges is a cloud-based platform that enables rapid and collaborative analysis of datasets in concert with other forms of biomedical data by utilizing High Throughput Sequencing (HTS) technologies. To interpret specifications, workflow engines like Reproducible Analyses for Bioinformatics (Rabix) Executor enable reproducibility by making data processing easier. Rabix, an open-source CWL executor, is embedded within the platform and orchestrates multi-instance and parallelizable execution on AWS and Google (http://rabix.org). The Rabix Composer, an integrated development environment for CWL, allows workflows to be constructed and executed locally and readily deployed on the platform, furthering interoperability. Seven Bridges Core Infrastructure enables standardized data analysis and collaboration support, as exemplified by Cavatica. Cavatica allows physicians to share and analyze genomic profiles of pediatric brain tumors when deciding on clinical treatment plans. Cavatica exemplifies the applications of reproducible data, allowing greater collaboration and treatment efficiency.

National Cancer Institute (NCI) Cloud Resources

The NCI Cloud Resources were formerly known as NCI Cancer Genomics Cloud (CGC) Pilots, which were conceptualized in 2013 to democratize access to NCI-generated genomic data and facilitate analysis. Three Cloud Pilot awardees – the Broad Institute (https://software.broadinstitute.org/firecloud/), the Institute for Systems Biology (http://cgc.systemsbiology.net/), and Seven Bridges (http://www.cancergenomicscloud.org/) have independently developed cloud-based analysis platforms. As a Software-as-a-Service built on commercial cloud architectures, these cloud resources offer the flexibility for researchers to utilize their own tools in the form of Docker containers. Tools can also be joined to form complex workflows described by Common Workflow Language (CWL) or Workflow Description Language (WDL). In a user-friendly graphical user interface, computation and data are encapsulated in a secured, access-controlled environment that also allows for sharing with collaborators.

Internet2 Community

Internet2, the U.S. research and education network, connects academic, government (including NIH, FDA and CDC), and life sciences companies. Internet2 also extends connectivity to the local level, including many healthcare institutions, through its high bandwidth U.S. Unified Community Anchor Network (U.S. UCAN) Program. Together these members constitute a diverse problem solving community that can share data frictionlessly at high speeds. Finally, over six million users at member institutions collaborate using Internet2's InCommon trust and federated identity management system. This enables virtualization of compute and storage resources, both private and cloud, to reduce costs and speed both information sharing and discovery. As this virtual infrastructure becomes more intelligently responsive to data driven operations, the BCO initiative promises to improve data findability and execution of distributed workflows through enhanced structuring of data.