Using Semantic Workflows to Disseminate Best Practices and Accelerate Discoveries in Multi-Omic Data Analysis

Yolanda Gil
Information Sciences Institute &
Department of Computer Science
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
gil@isi.edu

Shannon McWeeney
Division of Bioinformatics and Computational Biology
Department of Medical Informatics and Clinical Epidemiology
OHSU Knight Cancer Institute
Oregon Health and Science University
Portland, OR 97239
mcweeney@ohsu.edu

Christopher E. Mason
Department of Physiology and Biophysics &
Institute for Computational Biomedicine
Weill Cornell Medical College
Cornell University
1305 York Avenue
New York, NY 10021
chm2042@med.cornell.edu

Abstract

The goal of our work is to enable omics analysis to be easily contextualized and interpreted for development of clinical decision aids and integration with Electronic Health Records (EHRs). We are developing a framework where common omics analysis methods are easy to reuse, analytic results are reproducible, and validation is enforced by the system based on characteristics of the data at hand. Our approach uses semantic workflows to capture multi-step omic analysis methods and annotate them with constraints that express appropriate use for algorithms and types of data. This paper describes our initial work to use semantic workflows to disseminate best practices, ensure valid use of analytic methods, and enable reproducibility of omics analyses. Key elements of this framework are that it is knowledge-rich with regard to parameters and constraints that impact the analyses, proactive in the use of this knowledge to guide users to validate and correct their analyses and dynamic/adaptive as data sets evolve and change, all features that are critical for successful integration of omics analyses in a clinical setting.

Introduction

The advent of patient care that is guided by genomics, epigenomics, proteomics and metabolomics and other so-called 'omics' data types represents a new challenge for health informatics. Genomics based analyses in particular marks a break-through in the application of genetic testing for clinical decision-making. To handle these data, a wealth of complex statistical techniques and algorithms has been developed to process, transform, and integrate data. In addition, as new analytical tools and methods are constantly appearing in the field, translational scientists, clinicians and diagnostics laboratories are finding it increasingly challenging to keep up with the literature to select and evaluate among various state-of-the-art methods to analyze their data, as well as understand the implications for different algorithms on clinical diagnostics and biological research. There is a critical need for dynamic systems that can reanalyze and reinterpret stored raw data as knowledge evolves, and can incorporate genomic clinical decision support. This will enable dissemination of best practices and reproducibility of the analysis pipelines.

The goal of our work is to enable any lab to easily use omic analyses on one set of samples, have confidence in the results, re-execute the method seamlessly as new data becomes available, allow easy replication of biological results, and carry out meaningful comparisons across datasets for individual samples, clinical cohorts and populations.

We are developing a framework centered on *semantic workflows* to provide guidance and assistance for multi-omic data analysis, whereby the data is examined in each step of the process and relevant suggestions direct the analysis towards next options. Workflows have been used to manage complex scientific applications [Taylor et al 2007]. Workflows capture an end-to-end analysis composed of individual analytic steps as a dependency graph that indicates dataflow as well as control flow among steps. In the WINGS workflow system, we have extended workflows with semantic representations that support automatic constraint propagation and reasoning algorithms to manage constraints among the individual workflow steps [Gil et al 2011].

This paper how a library of pre-defined workflows that reflects best practices in clinical omics can facilitate reproducibility and standardization. We illustrate the use of semantic workflows to replicate two published studies that had taken months to perform and used proprietary software, while our replications took minutes by re-using generalized workflows built with open source software.

Reproducibility, Standardization, and Validation in Clinical Omics

The scale of omic data from biomedical and clinical researchers has recently expanded to an unprecedented level - from basic biology to translational medicine, multi-

omic data can enable phenomenal discoveries [Shendure and Ji 2008]. There is a dramatic shift from discovery research into clinical implementation. However, the ability to integrate and interrogate multiple 'omic data sets is critical for the understanding of disease and will only be accomplished through stringent data management, analysis, interpretation, and quantification. Ultimately, placing validated analytical tools in the hands of biomedical experts, and translating insights found between diverse datasets, will ensure that patients receive the correct diagnosis and individualized treatment.

Genetic testing for patient care has evolved tremendously over the past 50 years. Metaphase karyotyping has been used to diagnose disease since approximately 1960. The development of fluorescently or radioactively labeled probe hybridization approaches brought further advances in chromosomal analysis [Tsuchiya 2011]. The advent of PCR and the development of DNA sequencing allowed the first gene variant tests to be introduced into clinical laboratories. Subsequently, microarray technologies have supported genome-wide analyses of chromosomal gains and losses as well as global gene expression profiles. Deep sequencing technologies represent the next step forward and have the potential to replace several of these other diagnostic approaches. To handle these data, a wealth of complex statistical techniques and algorithms has been developed to process, transform, and integrate data. In addition, there is a further layer of complexity due to the evolving annotation for omics results that provides critical context for clinical interpretability.

A continually updating system that can aid in reproducibility, standardization, and validation of these workflows is needed to facilitate this transition from the research setting to the clinical setting.

Capturing Omics Analyses as Workflows

A computational experiment specifies how selected datasets are to be processed by a series of software or analytical components in a particular configuration. For example, biologists use computational experiments for analysis of RNA-seq or molecular interaction networks and pathways Computational workflows represent complex applications as a dependency graph of computations linked through control or data flow. Workflow systems manage the execution of these complex computations, record provenance of how results are generated, and allow end users with little or no programming background to create applications by reusing pre-defined workflows that others have built [Taylor et al 2007]. Popular workflow systems in omics include GenePattern [Reich et al 2006], Galaxy [Giardine et al 2005] and Taverna [Oinn et al 2006].

All these systems help users by providing easy access to heterogeneous tools organized in workflows. However, an important challenge for these systems is that they often exist with simple descriptions that provide no semantics as to what inputs they expect and what outputs they produce. Researchers often need better guidance and dynamic checks of the data, to ensure workflow validity and reproducibility. This is particularly critical as omics moves from research to clinical settings, where standardization of omics workflows is necessary to allow meaningful comparisons across datasets as well as updating results when new algorithms or data become available.

Reproducibility

Scientific articles often describe computational methods informally, often requiring a significant effort from others to reproduce and to reuse. Reproducibility is a cornerstone of scientific method, so it is important that reproducibility be possible not just in principle but in practice in terms of time and effort to the original team and to the reproducers. The reproducibility process can be so costly that it has been referred to as “forensic” research [Baggerly and Coombes 2009]. Studies have shown that reproducibility is often not achievable from the article itself [Bell et al 2009; Ioannidis et al 2009]. Retractions of publications do occur, more often than is desirable - a recent editorial proposed tracking the “retraction index” of journals to indicate the proportion of published articles that are later found problematic [Fang and Casadevall 2011]. Publishers themselves are asking the community to end “black box” science that cannot be easily reproduced [Nature 2006].

The need for reproducibility in the clinical arena is well documented. Clinical trials based on erroneous results pose significant threats to patients [Hutson 2010]. In addition, pharmaceutical companies have reported millions of dollars in losses due to irreproducible results that seemed initially promising [Naik 2011].

Computational reproducibility is in itself a relatively modern concept. Scientific publications could be extended so that they incorporate computational workflows, as many already include data [Bourne 2010]. However, without access to the source codes for the papers, reproducibility has been shown elusive [Hothorn and Leisch 2011]. Some systems exist that augment publications with scripts or workflows, such as Weaver for Latex [Falcon 2007] and GenePattern for MS Word [Mesirov 2010].

Repositories of shared workflows enable scientists to reuse workflows published by others and facilitate reproducibility [De Roure et al 2009]. The Open Provenance Model (OPM) [Moreau et al 2011] was developed to allow workflow systems to publish and exchange workflow execution provenance, therefore facilitating reproducibility.

Semantic Workflows in WINGS

WINGS is a semantic workflow system that assists scientists with the design and reproducibility of computational experiments [Gil et al 2011a; Gil et al 2011b]. Relevant publications and open source software are available from http://www.wings-workflows.org.

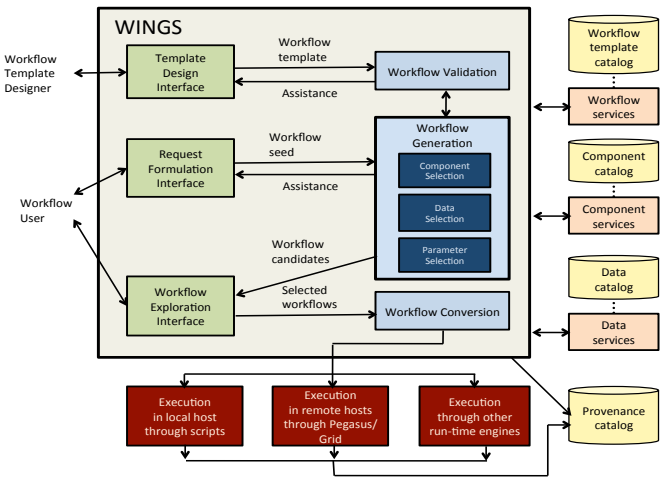

Figure 1. High-level architecture diagram of the WINGS semantic workflow system.

A high-level diagram of the architecture of WINGS is provided in Figure 1. A unique feature of WINGS is that its workflow representations incorporate semantic constraints about datasets and workflow components. WINGS represents semantic constraints that capture dataset properties and component requirements. Wings includes algorithms that use these representations for automated workflow elaboration, workflow matching, provenance and metadata generation, data-driven adaptive workflow customization, parallel data processing, workflow validation, and interactive assistance.

WINGS adopts the emerging W3C PROV standard for Web provenance to publish workflow executions [Gil and Miles 2013]. In addition, WINGS publishes workflow templates to enable reuse and reproducibility by other workflow systems [Garijo and Gil 2011]. The workflows and their execution records are published as semantic web objects using linked data principles, which means that all the provenance entities are accessible as web objects with a unique URI and represented in RDF. This includes the workflow execution and all associated artifacts such as data products, application codes, and parameter settings. An advantage of this approach is that the workflows can be linked to millions of other entities already published as linked data, including numerous biomedical data repositories (e.g., GO, KEGG, PDB). The integration of such domain knowledge with workflows has been explored in the SADI framework [Wood et al 2012].

Capturing Best Practices in Omics Analysis

We are developing a growing collection of workflows for genomic analysis, including population studies, inter- and intra-family studies, and next generation sequencing [Gil et al 2012]. It currently has workflows for: 1) ***Association tests***, including association test conditional on matching that includes population stratification, a general association test that assumes that outliers have been removed, and a structured association test; 2) ***Copy number variation (CNV) detection***, which use ensembles of algorithms with different algorithm combinations; 3) ***Transmission disequilibrium test (TDT)*** to conduct association testing for disease traits, in some cases incorporating parental phenotype information; and 4) ***Variant discovery*** from resequencing of genomic DNA and RNA sequencing.

The workflows include software components from the following packages: Plink (genome association studies toolset), R (statistical computing and graphics), PennCNV and Gnosis (CNV detection), Allegro and FastLink (linkage analysis), Burrows-Wheeler Aligner and SAMTools (sequence alignment), and Structure (population studies).

Figure 2 shows an RNA-Sequencing workflow, where beige boxes are data and blue circles are algorithms executed as the workflow runs. There are seven alignment steps (genome, junction, fusion, polyA, polyT, miR, and paired) in the analysis of RNA-Seq data, and each can be integrated with other workflows downstream. The yellow boxes highlight how the workflow is automatically elaborated by WINGS to process data collections in parallel, annotating each result with semantic metadata.

Semantic constraints are used to validate omic analyses by checking the integrity of the data. This quality control can reveal data fidelity and sample integrity problems, ruling out incorrect samples for many reasons (tube-mixups, non-paternity, technician error, labeling errors, etc). The workflows accept the familial relationships if known (pedigree file, or pedfile) and the raw genotyping data files (Affymetrix or Illumina) or sequencing (Illumina), import the data into memory, and format the data. For data with family information, we make extensive use of Plink tools as well as our own data-checking code. When pedigree data is present, it is used to validate the family structure and each sample's characteristics by examining all of the family members, their relationships, sex, and status (affected or unaffected). We examine the family structure and ensured that all of the family members are related, based on their pairwise identity-by-state distances. Then, we determine the "molecular sex" for each sample, based upon the rate of heterozygosity on the X chromosome. In each case, the workflow reasoners ensure the appropriate parameters and constraints are met before the next processing step is executed.

Reproducibility through Workflow and Provenance Publication

To illustrate how workflows can enable standardization of analysis and reproducibility, we discuss the replication of two published disease studies from the literature. The original studies did not use a workflow system, and all the software components were executed by hand. While these examples are focused on translational research in complex diseases, the framework is equally applicable to diagnostic clinical omics workflows as well.

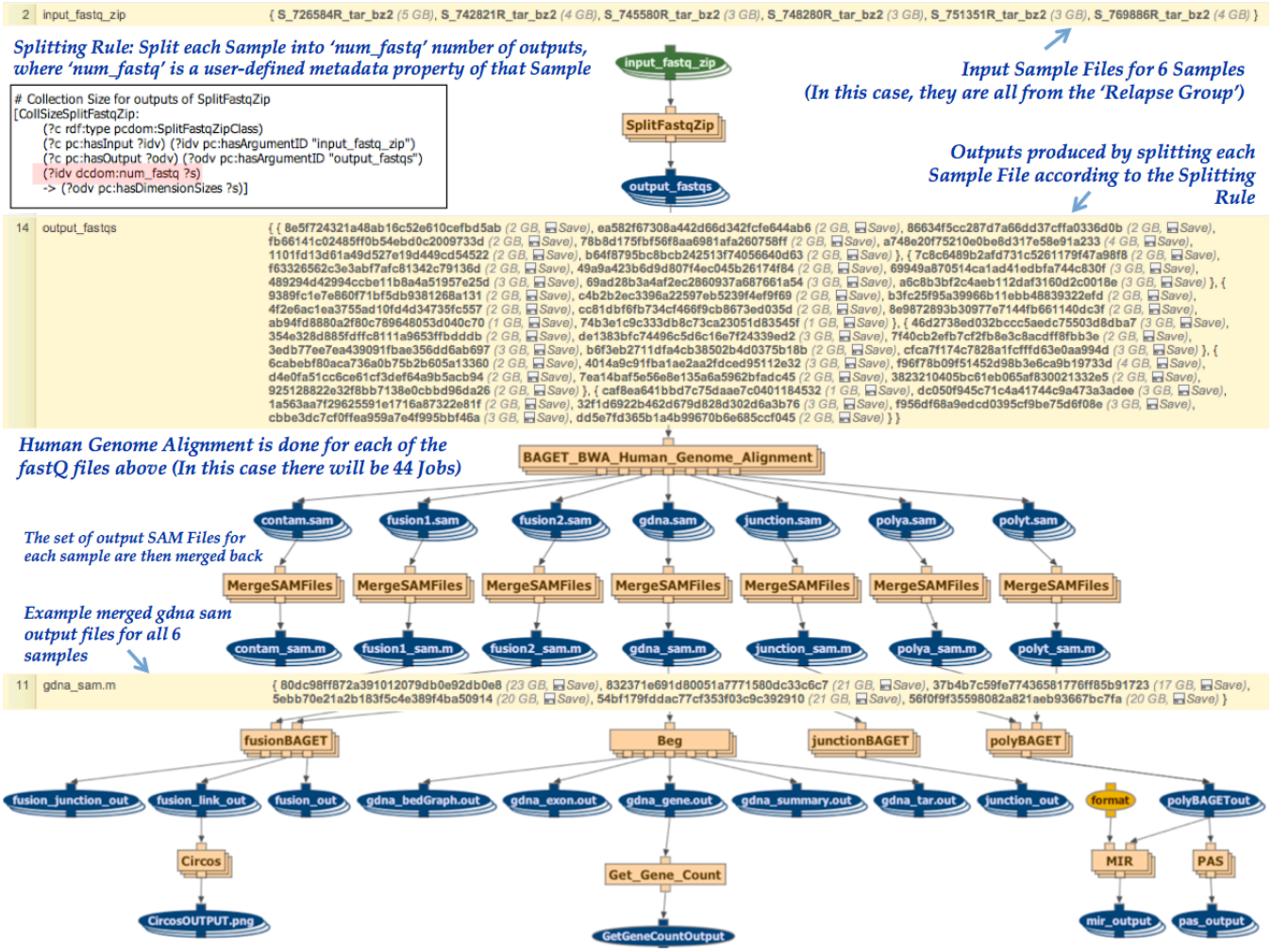

Figure 2: An RNA-Sequencing workflow is automatically elaborated by the system to process data collections in parallel.

We replicated the results of a study reported in [Duerr et al 06], which found a significant association between the IL23R gene on chromosome 1p31 and Crohn's disease. To reproduce the result, we used one of the workflows in our library for association test conditional on matching. It uses the Cochran-Mantel-Haenszel (CMH) association statistic to do an association test conditional on the matching done in the population stratification step. There are two input datasets to this workflow. One is a pedigree file, with one entry per individual with its unique identifier, gender, identifiers for each parent, phenotype information (a quantitative trait or an affection status), and optionally genotype information (given as two alleles per marker). The other input is a genetic map file that specifies the markers, one per line, including the chromosome identifier, SNP identifier, and position. The workflow components include the Inheritance-by-Structure (IBS) clustering algorithm from Plink for population stratification, the CMH association test from Plink, and the R package for plotting results. The pedigree dataset included 883 families with genotypic information for both parents and at least one offspring. The size of the file was 2.4 GB, the map dataset 10 MB. Figure 3 shows the results, each point plotted is a SNP and we highlight with a circle the points in Chromosome 1 with a log p value above 4.00. All five SNPs are in the IL23R gene, which was the main result of the study. The run time of the workflow was 19.3 hours. Most workflow components took minutes to execute. The population stratification step took 19 hours to run, as did the visualization step, both were executed concurrently.

We also replicated another previously published result for CNV association [Bayrakli et al., 2007]. Using one of our workflows for CNV detection, we saw a CNV at the expected locus over the PARK2 gene for early-onset Parkinson's disease (chr6:146,350,000-146,520,00). Figure 4shows the results. Each spot represents a probe on the arrays, with the x-axis representing coordinates on chromosome6 and the y-axis representing the average log2ratio of the patient versus control intensity at the same probe. Our workflow had 16 steps and run in 34 mins.

Some observations that we make from these studies are:

- *A library of carefully crafted workflows of select state-of-the-art methods will cover a very large range of genomic analyses.* The workflows that we used to

replicate the results were independently developed and were unchanged. They were designed with no notion of the original studies.

- *Workflow systems enable efficient set up of analyses*. The replication studies took seconds to set up. There was no overhead incurred in downloading or setting up software tools, reading documentation, or typing commands to execute each step of the analysis.
- *It is important to abstract the conceptual analysis being carried out away from the details of the execution environment*. The software components used in the original studies were not the same than those in our workflows. In the original study for Crohn's disease, the CMH statistic was done with the R package, and the rest of the steps were done with R and the FBAT software, while our workflow used CMH and the association test from Plink and the plotting from R. Our workflows are described in an abstract fashion, independent of the specific software components executed. In the original Parkinson's study, the Circular Binary Segment (CBS) algorithm for CNV detection was used, while our workflow used a state-of-the-art method that combines evidence from three newer algorithms. Our workflows contain state-of-the-art methods that can be readily applied.
- *Semantic constraints can be added to workflows to avoid analysis errors*. The first workflow that we submitted with the original Crohn's disease study dataset failed. Examining the trace and the documentation of the software for association test we realized that no duplicate individuals can be present. Upon manual examination we discovered that there were three duplicated individuals in the dataset. We removed them by hand and the workflow executed with no problems. The workflow now includes a constraint that the input data for the association test cannot contain duplicate individuals, which results in the prior step having a parameter set to remove duplicates. The time savings to future users could be significant.

Our framework allows easy replication of biological results and meaningful comparisons across datasets for individual samples, clinical cohorts and populations. Our goal is to enable omics analysis to be easily contextualized among appropriately matched, similarly analyzed, and biologically relevant data from within the same person (tumor/normal) or in the context of other genomes.

Discussion

In future work, we plan to further extend the means to filter, validate, and combine data that goes into the workflows. The integration and processing of the raw data is so critical because any errors in this first phase will contaminate the results of all subsequent steps. This requires adding to the workflows constraints for checking various aspects of the data such as sample integrity and genomic integrity. Additional steps and semantic

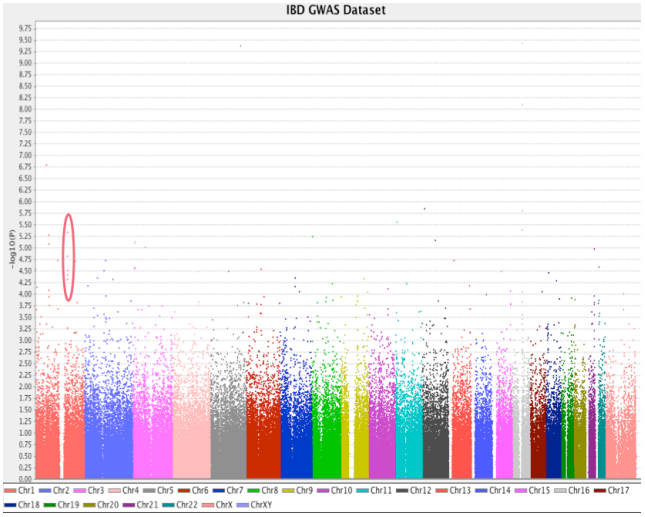

Figure 3. Results of the Crohn's disease replication.

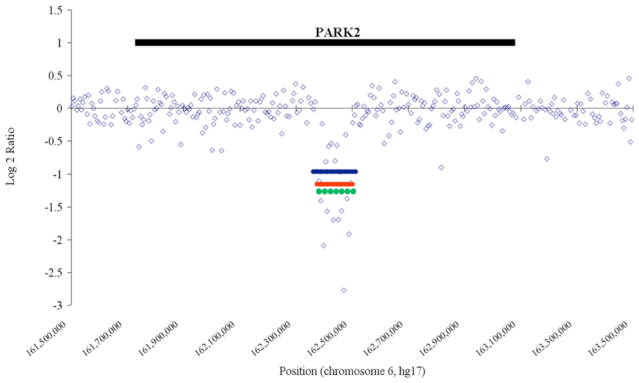

Figure 4. Results of the Parkinson's disease replication.

constraints can be added to expand the validity checks for gender, genome build, base composition, alignment rates, genome coordinates, and base types.

Figure 5 shows an example of an integrated RNA-seq workflow in WINGS that requires more advanced constraints and filtering. Alignments across N reference genomes or transcriptomes can be stored in BioHDF (http://www.hdfgroup.org/projects/biohdf/), and the reads can be hierarchically extracted for subsequent analysis. First, reads that map to a reference are used for variant prediction and allele-specific variation detection. Those reads are counted and converted into expression measures for all known genes/transcripts, which can then be used to predict the tissue source, then the remainder of the reads are used for gene fusion detection. The last set of reads are checked for the presence of any other genomes that are present, such as the re-constituted HPV genomes that we spiked into these sequences. If no other genomes are found, the system will prompt the user to attempt a de novo assembly of the remaining reads.

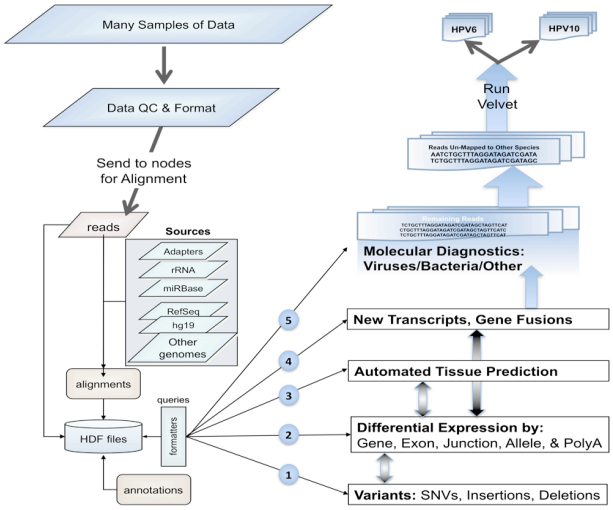

Figure 5. Integrated workflows for clinical omics.

Conclusions

We have described the use of semantic workflows to assist users in: 1) validating their use of workflows for complex genomic analyses, 2) readily and reliably replicating prior studies, and 3) finding published datasets to support their ongoing analysis. Semantic workflows can provide key capabilities needed to handle the complexities of population genomics, particularly given the new challenges of the upcoming NGS technologies.

Key elements of our approach that are critical for the effectiveness of our framework are that: 1) it is knowledge-rich with regard to parameters and constraints that impact the analyses, 2) it is proactive in the use of this knowledge to guide users to validate and correct their analyses, and 3) it is dynamic/adaptive as data sets evolve and change.

Our goal is to support reproducibility, standardization, and validation of omics workflows to accelerate the discovery of new genetic variations that contribute to human disease, as well as support the translation of these findings to the clinical and diagnostics setting.

Acknowledgments. We would like to thank Ewa Deelman and Varun Ratnakar for valuable discussions.

References

Baggerly, K. A. and Coombes, K. R. "Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-Throughput Biology." Annals of Applied Statistics, 3(4), 2009.

Bayrakli F, Bilguvar K, Mason, CE, et al. "Rapid identification of disease-causing mutations using copy number analysis within linkage intervals." Human Mutation, 28(12), 2007.

Bell AW, and the Human Proteome Organization (HUPO) Test Sample Working Group. "A HUPO test sample study reveals common problems in mass spectrometry–based proteomics." Nature Methods, 6(6), 2009.

Bourne, P. "What Do I Want from the Publisher of the Future?" PLoS Computational Biology, 2010.

De Roure, D; Goble, C.; Stevens, R. "The design and realizations of the myExperiment Virtual Research Environment for social sharing of workflows". Future Generation Computer Systems, 25, 2009.

Duerr RH, Taylor KD, et al. "A genome-wide association study identifies IL23R as an inflammatory bowel disease gene." *Science*, 314(5804):1461-3. Dec 1, 2006.

Falcon, S. "Caching code chunks in dynamic documents: The weaver package." Computational Statistics, (24)2, 2007.

Fang, C.F., and Casadevall, A. "Retracted Science and the retracted index". Infection and Immunity. 2011.

Garijo, D. and Y. Gil. "A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data." Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science (WORKS-11), Seattle, WA, 2011.

Giardine B, Riemer C, et al. "Galaxy: a platform for interactive large-scale genome analysis." *Genome Research* 15(10), 2005.

Gil, Y.; Gonzalez-Calero, P. A.; Kim, J.; Moody, J.; and Ratnakar, V. "A Semantic Framework for Automatic Generation of Computational Workflows Using Distributed Data and Component Catalogs." Journal of Experimental and Theoretical Artificial Intelligence, 23(4), 2011.

Gil, Y., Ratnakar, V., Kim, J., Gonzalez-Calero, P. A., Groth, P., Moody, J., and E. Deelman. "Wings: Intelligent Workflow-Based Design of Computational Experiments." IEEE Intelligent Systems, 26(1), 2011.

Gil, Y., Deelman, E. and C. Mason. "Using Semantic Workflows for Genome-Scale Analysis." International Conference on Intelligent Systems for Molecular Biology (ISMB), 2012.

Gil, Y. and S. Miles (Eds). "PROV Model Primer." Technical Report of the World Wide Web Consortium (W3C), 2013.

Hothorn, T. and F. Leisch. "Case Studies in Reproducibility." Briefings in Bioinformatics, 12(3), 2011.

Hull, D, Wolstencroft, K, Stevens, R, Goble, C, Pocock, M, Li, P, and T. Oinn. "Taverna: A Tool for Building and Running Workflows of Services", *Nucleic Acids Research*, 34, 2006.

Hutson, S. "Data Handling Errors Spur Debate Over Clinical Trial," Nature Medicine, 16(6), 2010.

Mesirov, JP. "Accessible Reproducible Research." Science, 327:415, 2010.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., et al. "The Open Provenance Model Core Specification (v1.1)." Future Generation Computer Systems, 27(6), 2011.

Nature Editorial. "Illuminating the Black Box." Nature, 442(7098), 2006.

Naik, G. "Scientists' Elusive Goal: Reproducing Study Results." The Wall Street Journal, December 2, 2011.

Reich, M., Liefeld, et al. "GenePattern 2.0". *Nature Genetics* 38(5):500-501, 2006.

Shendure, J & Ji, H. "Next-generation DNA sequencing." Nat Biotechnol 26, 1135-45, 2008.

Taylor, I., Deelman, E., Gannon, D., Shields, M., (Eds). *Workflows for e-Science*, Springer Verlag, 2007.

Tsuchiya, KD. "Fluorescence in situ hybridization." Clin Lab Med 31, 2011.

Wood, I., Vandervalk, B., McCarthy, L., and M. D. Wilkinson. "OWL-DL Domain-Models as Abstract Workflows". Proceedings of ISoLA, 2012.