Technologies for Genomic Medicine

The GMW, A Genetic Medical Workflow Engine

A RENCI TECHNICAL REPORT
TR-14-02

renci

CONTACT INFORMATION
Phillips Owen
Telephone: 919.445.9612
Email: powen@renci.org

List of Technical Terms and Websites

1000 Genomes Project, www.1000genomes.org

AnnoBot (Annotation Bot), www.renci.org/TR-14-04

Apache™ ActiveMQ STOMP – JMS mapping (Simple/Streaming Text Orientated Messaging Protocol – Java Mapping Services), activemq.apache.org/stomp.html

ApacheTM SOAP MTOM (Simple Object Access Protocol Message Transmission Optimization Mechanism), cxf.apache.org/docs/mtom.html

Apache™ SVN (Subversion)® Repository, subversion.apache.org

CANVAS (CAroliNa Variant Annotation System), www.renci.org/TR-14-04

CASAVA (Consensus Assessment of Sequence and VAriation), www.illumina.com/software/genome_analyzer_software.ilmn

Chrome development tools, www.google.com/intl/en/chrome/browser

CLIA (Clinical Laboratory Improvements Amendments), www.fda.gov/medicaldevices/deviceregulationandguidance/ivdregulatoryassistance/ucm124105.htm

ClinVar (Clinical Variants Resource database), www.ncbi.nlm.nih.gov/clinvar

daemons, en.wikipedia.org/wiki/Daemon_%28computing%29

dbSNP (Single Nucleotide Polymorphism Database), www.ncbi.nlm.nih.gov/SNP

Eclipse IDE (Integrated Development Environment), www.eclipse.org

ELSI (Ethical, Legal, and Social Implications) Research Program, www.genome.gov/elsi

ESP (Exome Sequencing Project), evs.gs.washington.edu/EVS

Firefox FireBug 1.10.3, getfirebug.com

GMW (Genetic Medical Workflow) Engine

HGNC (HUGO Gene Nomenclature Committee), www.genenames.org

HGMD® (Human Gene Mutation Database), www.hgmd.cf.ac.uk/ac/index.php

iRODS (integrated Rule-Oriented Data System), www.irods.org/index.php/IRODS:Data_Grids,_Digital_Libraries,_Persistent_Archives,_and_Real-time_Data_Systems

JQuery 1.7.1, jquery.com

JQWidgets (JQuery widgets), www.jqwidgets.com

MaPSeq (Massively Parallel Sequencing) System, www.renci.org/TR-14-03

Microsoft IIS 7.0 (Internet Information Services), www.iis.net

Microsoft SQL Server 2008 R2, www.microsoft.com/en-us/sqlserver/product-info.aspx

Microsoft SQL Server Management Studio, www.microsoft.com/en-us/download/details.aspx?id=8961

MySQL (Structured Query Language), www.mysql.com

OSG (Open Science Grid), www.opensciencegrid.org

PHP 5.3 (Hypertext Preprocessor), www.php.net/manual/en/intro-whatis.php

PostgreSQL database, www.postgresql.org

PostgreSQL pgAdmin, www.pgadmin.org

pythonTM modules, www.python.org

REDCapTM (Research Electronic Data Capture) application, www.project-redcap.org

RefSeq (Reference Sequence Collection), www.ncbi.nlm.nih.gov/refseq

Sparx Enterprise Architect, www.sparxsystems.com

SQL Server, www.microsoft.com/en-us/sqlserver/default.aspx

TeraGrid, info.teragrid.org

The Team*

Phillips Owen, RENCI Research Software Architect; **Stanley Ahalt**, PhD, RENCI Director and Professor, Department of Computer Science, UNC Chapel Hill; **Jonathan Berg** MD, PhD, Assistant Professor, Department of Genetics UNC Chapel Hill; **Joshua Coyle**, RENCI New Media Specialist; **James Evans**, MD, PhD, Professor, Departments of Genetics and Medicine, UNC Chapel Hill and Editor-in-Chief, *Genetics in Medicine*; **Karamarie Fecho**, PhD, Medical and Scientific Writer, RENCI; **Daniel Gillis**, RENCI Software Developer; **Charles P. Schmitt**, PhD, RENCI Chief Technical Officer and Director of Informatics and Data Science; **Dylan Young**, RENCI Software Developer; and **Kirk C. Wilhelmsen**, MD, PhD, RENCI Director of Biomedical Research, RENCI Chief Domain Scientist for Genomics, and Professor, Departments of Genetics and Neurology, UNC Chapel Hill.

TR-14-02 in the RENCI Technical Report Series, April 2014
Renaissance Computing Institute
University of North Carolina at Chapel Hill
919-445-9640
www.renci.org

How to reference this paper

Owen, P., Ahalt, S., Berg, J., Coyle, J., Evans, J., Fecho, K., Gillis, D., Schmitt, C., Young, D. & Wilhelmsen, K. (2014): Technologies for Genomic Medicine: The GMW, A Genetic Medical Workflow Engine. RENCI, University of North Carolina at Chapel Hill. http://dx.doi.org/10.7921/G0KW5CXC

About RENCI

RENCI is an institute of the University of North Carolina at Chapel Hill that develops and deploys advanced technologies to enable research discoveries and practical innovations. The institute was launched in 2004 as a collaborative effort involving UNC Chapel Hill, Duke University, and North Carolina State University. For more information, see www.renci.org.

*Phillips Owen serves as technical lead on the GMW Engine; Kirk Wilhelmsen serves as Principle Investigator and Director of RENCI's Biomedical Research division, which is leading the development of the GMW Engine; all other team members are listed alphabetically.

Introduction

Genomic data are rapidly amassing as a result of recent advancements in next-generation genomic sequencing and other high-throughput "-omics" technologies (Mardis, 2008; Horvitz and Mitchell, 2010; Koboldt et al., 2010; Kahn, 2011). Yet, we are far from an era of routine genetic screening (Evans and Berg, 2014). In order to take full advantage of the wealth of genomic data available today, and thereby better serve patients, technological advances are required to enable the secure, cost-effective, efficient, and accurate processing of genome-wide data, from sample collection in the clinic to physician or researcher interpretation of results (Ahalt et al., 2014; the Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data, 2013; Data and Informatics Working Group, National Institutes of Health BD2K Initiative, 2012).

Herein, we describe the Genetic Medical Workflow (GMW) Engine—an open source system that provides end-to-end capture, analysis, validation, and reporting of genome-wide data for use in research and routine clinical care.

The GMW Engine

The GMW Engine was developed initially to support a National Institutes of Health (NIH)–funded clinical research study, "North Carolina Clinical Genomic Evaluation by NextGen Exome Sequencing" (NCGENES; Foreman et al., 2013) at the University of North Carolina at Chapel Hill (UNC). NCGENES has both clinical and research arms and aims to explore the use of whole exome sequencing data in genomic medicine.

The initial development of the GMW Engine was prompted by an early recognition that in order to achieve the goals of NCGENES, a comprehensive solution was required for the management of numerous people, processes, samples, and information—a complex endeavor. Initially, RENCI evaluated existing open source or proprietary workflow management systems; however, none of the existing systems were deemed capable (without major modification) of managing all of the disparate groups and legacy data systems in place at UNC. A custom solution was needed to meet the following high-level criteria:

- Present a secure user interface (UI) to capture and display contextually relevant information to and from users representing greater than 20 unique study roles;
- Manage and orchestrate complex processes that span numerous UNC laboratories and research teams;
- Orchestrate initial, secondary, and tertiary data analysis pipelines on multiple UNC compute clusters;
- Automatically collect analysis results and situational awareness information from multiple and disparate UNC data systems; and
- Monitor and audit user and process performance, as well as overall system health.

All of these features were incorporated into the custom-built GMW Engine. The GMW Engine serves as a centralized workflow manager; it executes discrete, automated- or user-driven workflows, UIs, and tracking systems (Figure 1). Specifically, it activates and tracks workflows related to: patient/subject flow from the initial clinic visit to consultation regarding genomic findings to follow-up visits; genetic sample flow from collection to processing to sequencing; and data flow from analysis to annotation to reporting. The GMW Engine provides several services via this process: system integration; system management; quality control; auditing; signaling; and reporting.

To understand the GMW Engine and the operations of the different workflows, consider the Project Operations workflow. This is where operations specific to a research project take place, from the identification of potential subjects to enrollment and informed consent to collection of blood for the processing of genomic DNA. The Project Operations workflow also involves interactions between the clinician researcher (or ELSI researcher) and the patient/subject.

Each step of the Project Operations is securely tracked by the GMW Engine such that only authorized persons (e.g., the researcher, research nurse, information technology staff) can view the status of the project at any given time. Automated tracking also allows for auditing and signaling to ensure compliance with all privacy, security, and ELSI requirements. It should be noted that the Project Operations workflow is comprised of more than one workflow, each of which is orchestrated by the GMW Engine. For example, the Initial Subject Enrollment sub-workflow (described under Use Case #2) is just one of several sub-workflows that are managed under the Project Operations workflow.

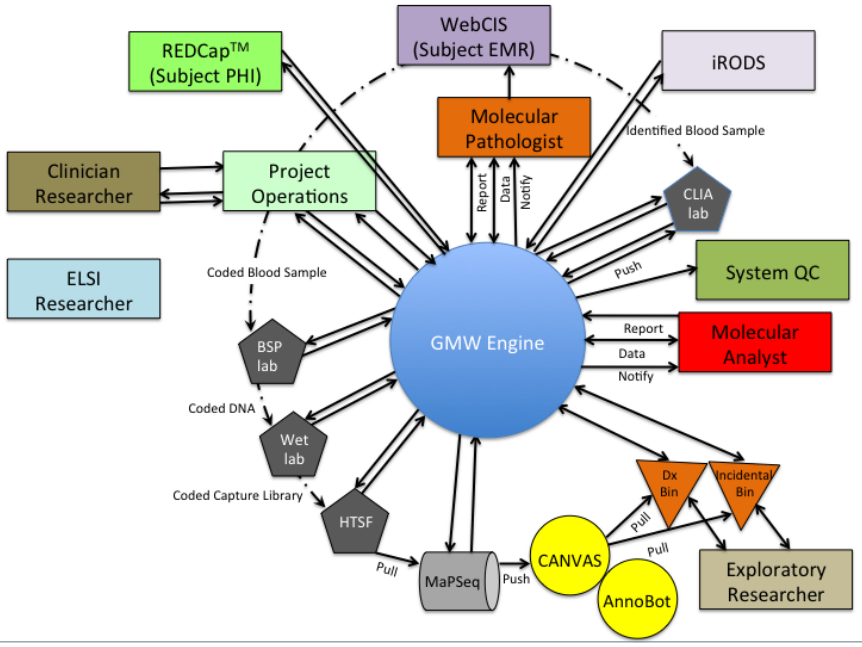

Figure 1. A schematic showing the workflows managed by the GMW Engine, with arrows depicting the flow of information. AnnoBot = Annotation Bot; BSP lab = BioSpecimen Processing laboratory; CANVAS = CAroliNa Variant Annotation Store; CLIA lab = a laboratory certified to meet U.S. Congressional Clinical Laboratory Improvements Amendments; Dx = diagnostic; ELSI Researcher = Ethical, Legal, and Social Implications Researcher; EMR = Electronic Medical Record; iRODS = integrated Rule-Oriented Data System; MaPSeq = Massively Parallel Sequencing system; PHI = Protected Health Information; QC = Quality Control; WebCIS = Web-based Clinical Information System; Wet lab = basic science laboratory.

Completion of the Project Operations workflow automatically leads, via the GMW Engine, to the processing of the coded blood sample by the BioSpecimen Processing (BSP) laboratory, where a new BSP Laboratory Information Management System (LIMS)–based workflow is initiated to track the initial processing of samples (i.e., DNA isolation). The BSP LIMS–based workflow is fully integrated with the GMW Engine, in order to keep the systems synchronized. Of note, all subject Protected Health Information (PHI) is securely stored using the open source REDCap™, which is also integrated with the GMW Engine. The PHI data are derived from the Web-based Clinical Information System (WebCIS), which is UNC Health Care's homegrown Electronic Medical Record (EMR) system. (We note that UNC Health Care is transitioning to the commercial Epic EMR system, so further modification of the GMW Engine is expected. The system is designed for flexibility, so modifications require minimal effort.)

After the BSP workflow has been successfully executed and coded DNA has been obtained, the samples are sent to a basic science laboratory in UNC's Genetic Medical Building, where a new workflow is initiated. There, samples undergo further processing (i.e., DNA amplification and other steps in preparation for sequencing). Completion of secondary sample processing leads to the execution of the High-Throughput Sequencing Facility (HTSF) workflow, where the raw genomic sequencing data are generated. Both the Genetic Medical Building and HTSF workflows are managed using the BSP LIMS, although all workflow steps are tracked by the GMW Engine for auditing purposes and to allow only authorized users to view the status of any given workflow.

Completion of the HTSF workflow leads to the execution of the MaPSeq system, which is designed to perform multiple levels of genomic data analysis on a massively parallel computational cluster (Reilly et al., 2014). Specifically, MaPSeq is an open source, plugin-based, service-oriented application developed by RENCI in collaboration with UNC's Information Technology Services, Research Computing Division, the UNC High-Throughput Sequencing Center, and Lineberger Comprehensive Cancer Center. MaPSeq provides a framework for facilitating the construction, deployment, and activation of project-specific, downstream, sequence analysis pipelines. The analysis pipelines invoke project-defined computation on the output from the raw HTSF data, such as genomic sequence alignment and variant calling. MaPSeq is designed to opportunistically take advantage of available institution-wide and cloud-based computational resources, including OSG, TeraGrid, and computational clusters available at RENCI and UNC's Department of Computer Science. MaPSeq was developed initially to support a genomics research project within the Lineberger Comprehensive Cancer Center at UNC, but it is now used to support numerous high-throughput sequencing projects at UNC, including NCGENES.

The MaPSeq workflow pushes data into CANVAS[1], which works together with AnnoBot as open source,

[1] CANVAS (CAroliNa Variant Annotation System) was originally termed VarDB (Variant DataBase)

homegrown technologies to enable the capture, storage, and updating of annotations to provide critical clinical interpretations of genomic data and metadata to attribute provenance or "ownership" and record the history of a given data set (e.g., type of sample, laboratory processing steps, analysis steps, validity and reliability estimates, etc.) (Bizon et al., 2014). CANVAS is a relational PostgreSQL database that stores up-to-date annotation and related metadata on genomic variants. As variant data from GMW Engine–supported research projects are pushed into CANVAS, they are matched against reference variant data from RefSeq and annotated accordingly. Additional annotation and associated metadata on variants are pulled into CANVAS by AnnoBot. AnnoBot is comprised of a set of python™ modules, as well as software driver code, designed to automatically monitor targeted databases for updates, extract new or revised annotation, and add that annotation to the variant data in CANVAS. The databases that are currently monitored by AnnoBot include dbSNP, the 1000 Genomes Project, ESP, HGNC, HGMD®, and ClinVar. CANVAS and AnnoBot together provide interpretations of genomic variant data that can be used to evaluate the diagnostic capability of identified genomic variants.

For NCGENES, CANVAS uses a Clinical Binning schema (ClinBin) to compute on the annotated variant data in order to determine which of two database Bins the identified patient/subject variants should get pushed into: the Diagnostic (Dx) Bin or the Incidental Bin. The Dx Bin includes variants that were targeted for a given patient/subject on the basis of a defined phenotype and have established clinical validity and utility (Shoenbill et al., 2014); in contrast, the Incidental Bin includes incidental findings[2], or variants that were identified during the sequencing effort but were not targeted as part of the diagnosis. (See Foreman et al., 2013 for a more detailed description of the binning process.) Note that only the targeted diagnostic findings are used for clinical care; incidental findings are used for research purposes only, unless they are classified as "medically actionable" under guidelines put forth by the American College of Medical Genetics and Genomics (Foreman et al., 2013; Green et al., 2013).

Table 1 shows the current number of genes/loci associated with the different diagnostic classes currently explored by NCGENES. Note that the data in both the Dx and Incidental Bins can be used for exploratory research (as opposed to the initial hypothesis-driven research), in which case the researcher re-analyzes the data post hoc to data-mine for unrecognized, potential associations between phenotype and genotype. Note also that the Incidental Bin is further subdivided on the basis of the degree of clinical validity and utility of

Table 1. Number of targeted genes associated with different diagnostic classes in the NCGENES study.

Number genes/loci	Diagnostic Class
31	Arrhythmia
15	Autoinflamation
82	Cancer
75	Cardiomyopathy
449	CNS
420	Dysmorphology
59	Immunodeficiency
521	Intellectual Disability and Autism
46	Leukodystrophy
69	Microcephaly
109	Mitochondrial
15	Myasthenia
99	Myopathy
315	Nueromuscular Disorders
80	Neuropathy
5	Polyposis
18	Progeria
214	Retina
46	Rhabdomyolysis
103	Seizure
162	Skeletal Dysplasia
45	Spastic Paraplegia
91	Storage Disorders
12	Thoracic Aneurysm/ Dissection

As required by the 1988 U.S. Congressional CLIA, patient (as opposed to research) samples are processed in a CLIA-certified laboratory to ensure analytical validity (Shoenbill et al., 2014) and to meet the quality standards put forth by the Centers for Medicare & Medicaid Services and the Food & Drug Administration. After processing in MaPSeq, variant data that are derived from a patient sample are reviewed by a Molecular Analyst, who determines which

[2]"Incidental findings" refer to genomic variants that are identified as a result of a genetic screening test but are unrelated to genes targeted by the testing. The ethical use of incidental findings is a topic of much debate (Evans and Berg, 2014).

of the identified mutation(s) is clinically significant. Those results get passed to a Molecular Pathologist, who performs a secondary sequence analysis of the genetic sample in order to ensure that the mutation(s) truly exists (i.e., to verify the genetic finding[s]). The Molecular Pathologist's final report is then sent to WebCIS for incorporation into the patient's EMR. Each step in these workflows is executed and tracked by the GMW Engine.

iRODS (Moore and Marciano, 2005; Rajasekar et al., 2010a,b; Schmitt et al. 2013) is used by the GMW Engine for secure data transfer and indexing among the disparate data analysis systems that are managed by the GMW Engine. iRODS is an open source, policy-based solution to access, share, integrate, publish, preserve, and manage data and associated metadata among remote data sources and diverse user communities. iRODS was developed by the Data Intensive Cyber Environments groups at UNC and the University of California at San Diego, with contributions from RENCI and other groups through the iRODS Consortium. iRODS was architected and designed to allow different adopter groups, with differing institutional goals and security concerns, to develop and deploy policies for data sharing that are specific to organizational needs. The GMW Engine relies on iRODS for secure, policy-based data transfer.

Finally, background daemons perform a continuous Quality Control (QC) check on the GMW Engine and the various systems and processes it relies on. The daemons use process connectors to query systems in order to track patients/subjects/samples/data and send error notification signals or alerts to Administrators and the staff member(s) who is responsible for the item of interest at that particular stage of processing. QC reports are also periodically generated for auditing.

Examples of GMW Engine Functionality

Although the GMW Engine was developed initially for NCGENES, it has since been modified and expanded for use in several additional research studies (see Impact section), and development continues as new user needs and tools become available. The workflows that are invoked by the GMW Engine are specific for each project and tailored to achieve the aims of that project. Each workflow depicted in Figure 1 is typically comprised of a comprehensive set of specific tasks organized in a decision tree or a linked subset of workflows organized in a similar manner.

We present two use cases for the GMW Engine: (1) the overall GMW Engine workflow processes and UIs engaged by NCGENES; and (2) the Initial Subject Enrollment and Genomic Sequencing workflows invoked by NCGENES.

Use Case #1: GMW Engine Workflow Processes and UIs for NCGENES

Figure 2 depicts the GMW Engine workflow processes that are engaged by NCGENES and specific to that project. (Not shown are the underlying REDCap™, iRODS, and System QC systems. Also not shown are the ELSI Researcher and Exploratory Researcher.) The process begins with step (1), when the Clinician Researcher activates the Project Operations workflow, which includes the Initial Subject Enrollment workflow discussed below. The numerical steps can then be traced to show the flow of patients, subjects, samples, and data. The final step, as outlined here, is step (20), in which the clinical report from the Molecular Pathologist is loaded into WebCIS for incorporation into the patient's EMR.

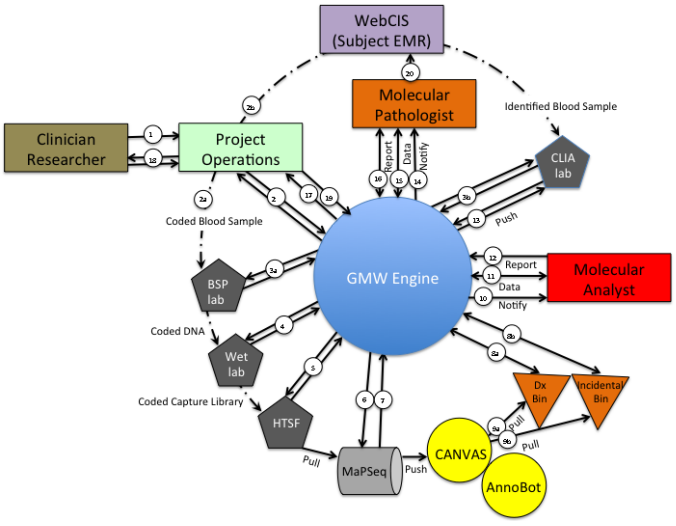

Figure 2. The main GMW Engine workflows engaged by NCGENES, with the flow of information marked as numerical steps. AnnoBot = Annotation Bot; BSP lab = BioSpecimen Processing laboratory; CANVAS = CAroliNa Variant Annotation Store; CLIA lab = a laboratory certified to meet U.S. Congressional Clinical Laboratory Improvements Amendments; Dx = diagnostic; EMR = Electronic Medical Record; MaPSeq = Massively Parallel Sequencing system; WebCIS = Web-based Clinical Information System; Wet lab = basic science laboratory.

A unique feature of NCGENES is its UIs. RENCI worked with NCGENES investigators to develop comprehensive UIs that are currently being used to support the NCGENES research project and will be evaluated for use as general Genomic Clinical Decision Support tools. Two example UIs are shown in Figures 3 and 4. The UI shown in Figure 3 displays study status and details for an individual patient or subject (identified in the figure as NCG_00256) and includes information related to diagnostic and incidental genomic findings, completed NCGENES workflows, current status (in terms of study completion), and whether the subject is in compliance with the study protocol. This UI provides information that is easy to read and interpret and can be used by any member of the study team, from Study Coordinator to Clinician Researcher to System Administrator.

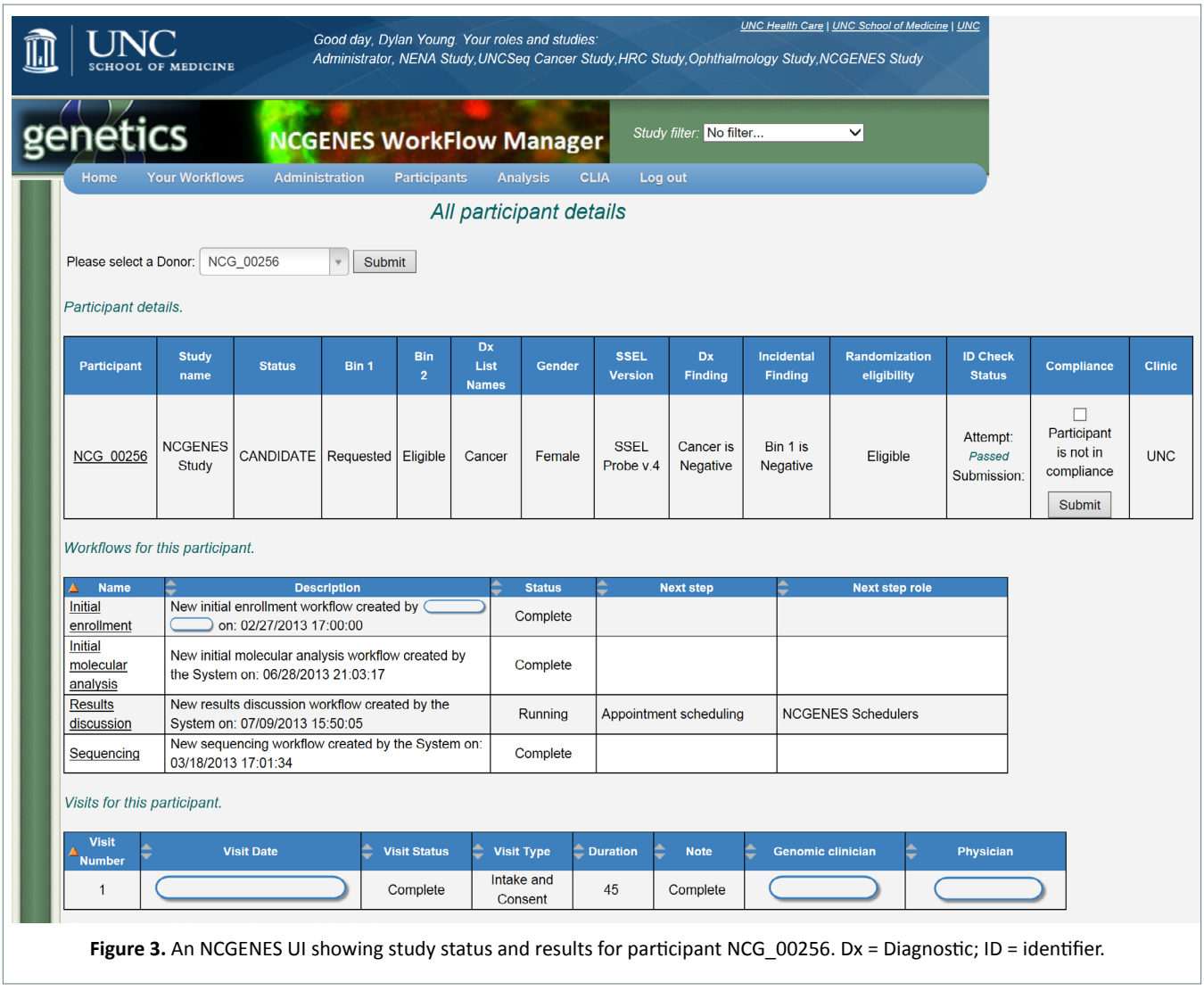

Figure 3. An NCGENES UI showing study status and results for participant NCG_00256. Dx = Diagnostic; ID = identifier.

In contrast, the UI shown in Figure 4 provides more comprehensive, detailed information than that shown in Figure 3. This UI was designed for use by the Molecular Analyst; it provides all of the information required to interpret the genomic sequencing results and reach a conclusion regarding an individual patient or subject. For example, information is provided on the effect of the variant on protein structure and function, the variant's accession number (if available), QC metrics, annotation derived from other sources, and molecular transcript information. Many of the UI fields contain hyperlinks to additional data sources, including the annotation sources that are monitored by AnnoBot and pushed back into CANVAS. The Molecular Analyst UI requires advanced training in the interpretation of fields and thus would not be used by a Study Coordinator, System Administrator, or any member of the study team other than the Molecular Analyst.

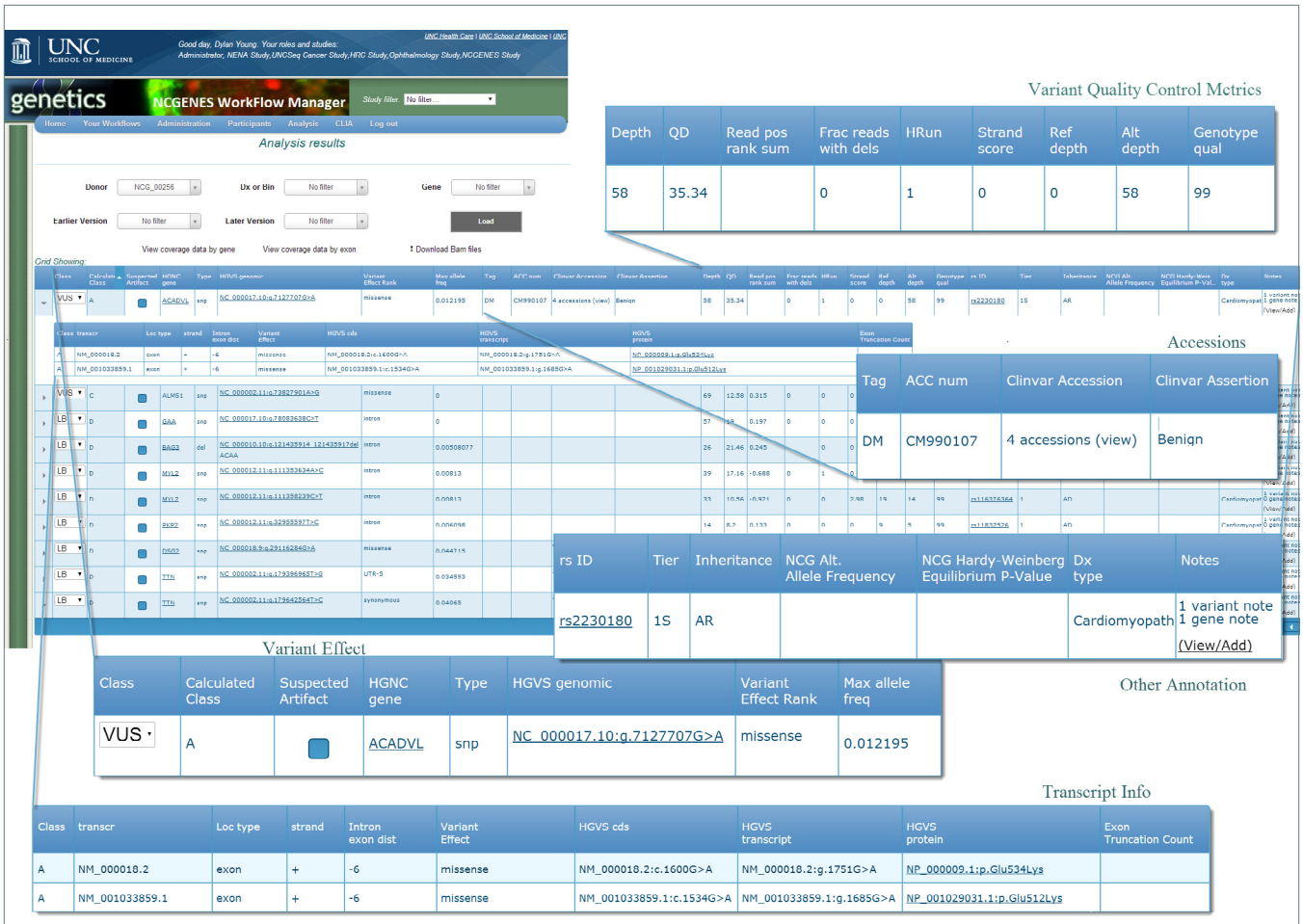

Figure 4. An NCGENES UI showing detailed results for review by the Molecular Analyst. Note that this UI is intentionally more comprehensive and detailed than the UI shown in Figure 3 because it is designed to provide all of the information required by the Molecular Analyst to analyze the results for a given patient or subject. The blow-ups show the types of information available through this UI.

Use Case #2: Workflow Schematics for NCGENES

As discussed, each of the workflows depicted in Figures 1 and 2 typically involves numerous steps and processes and often includes sub-workflows. One such sub-workflow, under Project Operations, is the Initial Subject Enrollment workflow (Figure 5). Note that each and every step in this seemingly "simple" workflow is specified and tracked by the GMW Engine. This level of detail provides for a comprehensive, secure process to facilitate genomic research.

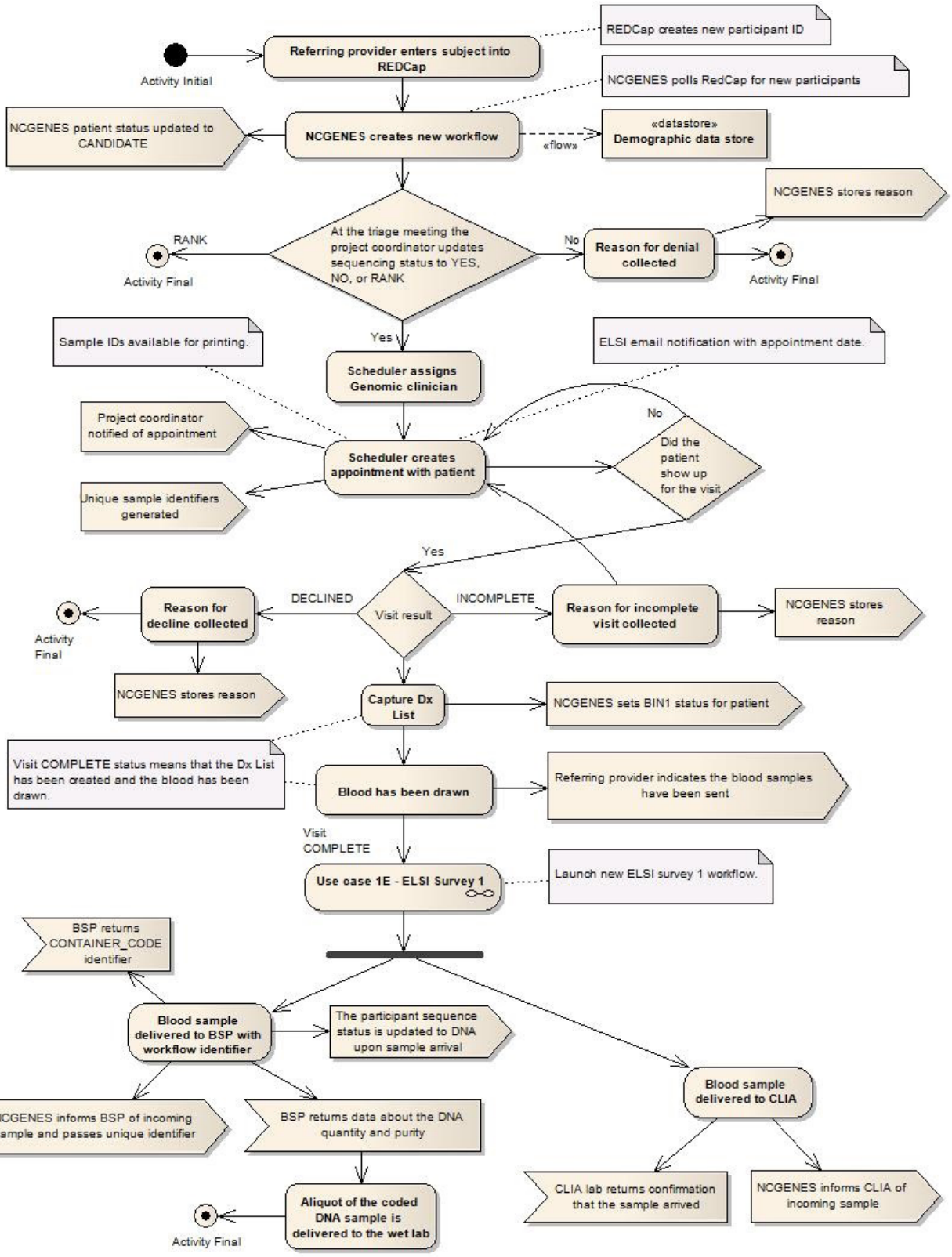

Figure 5. The Initial Subject Enrollment sub-workflow invoked during the execution of the Project Operations workflow. Note the complexity of the sub-workflow. The GMW Engine tracks each step of this sub-workflow and any others that are engaged by a given research project. BSP = BioSpecimen Processing laboratory; CLIA lab = a laboratory certified to meet U.S. Congressional Clinical Laboratory Improvements Amendments; Dx = diagnostic; IDs = identifiers; iRODS = integrated Rule-Oriented Data System; NCGENES = North Carolina Clinical Genomic Evaluation by NextGen Exome Sequencing; wet lab = basic science laboratory.

An important workflow is the Genomic Sequencing workflow (Figure 6). Note that this workflow contains its own sub-workflows, including the sequence analysis workflow used by MaPSeq and the binning workflow invoked by CANVAS. Of mention, communication and data transfer between the MaPSeq and CANVAS workflow pipelines are managed by iRODS. In particular, the MaPSeq workflow is registered with iRODS and uses iRODS to request a table in CANVAS, as needed. The GMW Engine is integrated with iRODS, MaPSeq, and CANVAS and manages the request by using metadata tags in iRODS to automatically look up the appropriate data files in MaPSeq and load those files into CANVAS.

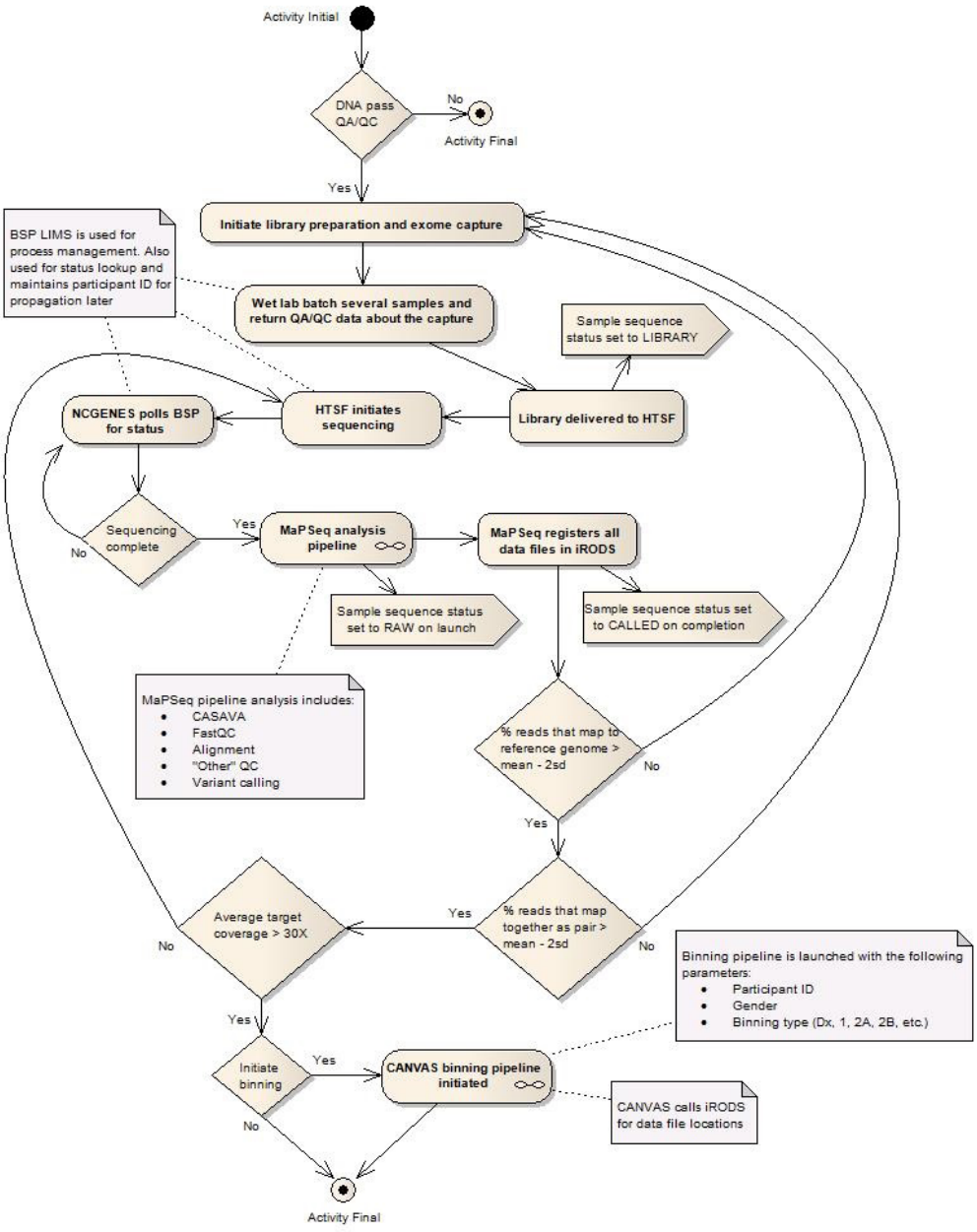

Figure 6. The Genomic Sequencing workflow. Note that this workflow invokes several sub-workflows, including the sequence analysis workflow used by MaPSeq and the binning workflow used by CANVAS. The GMW Engine tracks each step of the overall workflow and its sub-workflows. BSP = BioSpecimen Processing laboratory; CANVAS = CAroliNa Variant Annotation Store; CASAVA = Consensus Assessment of Sequence And VAriation; Dx = Diagnostic; HTSF = High-Throughput Sequencing Facility; ID = identifier; LIMS = Laboratory Information Management System; NCGENES = North Carolina Clinical Genomic Evaluation by NextGen Exome Sequencing; QA = Quality Assurance; QC = Quality Control; sd = standard deviation; vcf = variant call format.

Conclusion

The GMW Engine is an open source architecture that seamlessly coordinates numerous workflows, sub-workflows, samples, data, and people to provide an end-to-end approach to genomics, from initial clinic visit to reporting of genomic findings, thus enabling the secure and efficient use of whole-genome data in genomic research today and in genomic medicine in the near future.

Key Features:

- Architecture is open source.
- Numerous open source technologies are incorporated.
- UIs can be tailored to meet any user's needs Engine is modifiable, extendable, and scalable.
- Workflows are customizable.
- Workflows can be modified while running.
- Multiple workflows are capable of running simultaneously.

Underlying Software and Technologies:

Technology Stack:

- Apache™ SOAP MTOM
- Apache™ ActiveMQ STOMP – JMS mapping
- iRODS
- Microsoft IIS 7.0
- Microsoft SQL Server 2008 R2
- PHP 5.3
- JQuery 1.7.1
- JQWidgets
- Several database connectors, including SQL Server, MySQL, Oracle, and PostgreSQL
- Multiple UI plugins, including a calendar, barcodes, etc.

Development Environment:

- Apache™ SVN® Repository
- Chrome development tools
- Eclipse IDE
- Firefox FireBug 1.10.3
- Microsoft SQL Server Management Studio
- PostgreSQL pgAdmin
- Sparx Enterprise Architect

Impact:

- Currently supports variant annotation for the following research programs: (1) National Human Genome Research Institute–funded NCGENES, "North Carolina Clinical Genomic Evaluation by NextGen Exome Sequencing" (Dr. James Evans, PI), which is conducting whole exome sequencing of >2,000 patient samples drawn from multiple disease categories; (2) National Institute of Child Health and Development–funded NC Nexus, "North Carolina Newborn Exome Sequencing and Newborn Screening Disorders" (Dr. Cynthia Powell, PI), which aims to conduct whole exome sequencing on 400 patient samples; (3) UNCSeq, which applies tumor sequencing technology for >2,000 patient samples in order to identify mutations that are amenable to targeted treatments; and (4) National Institute on Drug Abuse–funded NIDASeq, "Deep Sequencing Studies for Cannabis and Stimulant Dependence" (Dr. Kirk Wilhelmsen, PI), which is conducting whole genome sequencing of ~5,500 patient samples.
- Also supports the NIH-funded Clinical Genome Resource (ClinGen) initiative (Dr. Jonathan Berg, Site PI), which involves a national effort to develop consensus annotation for the NIH Clinical Variant (ClinVar) database.
- Aggregates and stores ~6,000 additional genomes derived from public databases and used for analysis in ongoing genomic research studies; these are obtained from the 1000 Genomes project, The Cancer Genome Atlas project, the national Exome Sequencing Project, and Complete Genomics.

Acknowledgements

This project was conceptualized and funded by RENCI in collaboration with the North Carolina Translational and Clinical Sciences Institute and UNC's Information Technology Services Research Computing, with additional funding from the National Institutes of Health (1R01-DA030976-01, 1U01-HG006487-01, 5UL1-RR025747-03, 1U19-HD077632-01, 1U01-HG007437-01) and UNC's Lineberger Comprehensive Cancer Center.

The GMW Engine was awarded a 2013 HPC Innovation Excellence Award from the International Data Corporation and a 2013 Health IT Innovation Award from the North Carolina Healthcare Information and Communications Alliance.

Karen Green provided editorial and design support for the preparation of this technical report. Dr. Christopher Bizon provided assistance with the NCGENES screenshots.

References and Resources

1000 Genomes Project. (An international project designed to establish a public database containing detailed annotation on human genetic variation, both healthy and disease-related. Funded and maintained by the National Center for Biotechnology Information.) www.1000genomes.org. [Accessed December 13, 2013]

Ahalt S, Bizon C, Evans J, Erlich Y, Ginsberg G, Krishnamurthy A, Lange L, Maltbie D, Masys D, Schmitt C, Wilhelmsen K. Data to Discovery: Genomes to Health. A White Paper from the National Consortium for Data Science; 2014. RENCI, University of North Carolina at Chapel Hill. dx.doi.org/10.7921/G03X84K4. [Accessed Febuary 4, 2014]

AnnoBot (Annotation Bot), www.renci.org/TR-14-04 [Accessed March 27, 2014]

Apache™ SOAP MTOM (Simple Object Access Protocol Message Transmission Optimization Mechanism). (A Web-based mechanism to specify an optimized method for sending binary data as part of a SOAP message between a remote client and a database.) cxf.apache.org/docs/mtom.html. [Accessed January 30, 2014]

Apache™ ActiveMQ STOMP – JMS mapping (Simple/Streaming Text Orientated Messaging Protocol – Java Mapping Services). (A protocol to enable mapping of JMS messages over an open source message and integration patterns server.) activemq.apache.org/stomp.html. [Accessed January 30, 2014]

Apache™ SVN (Subversion)® Repository. (An open source version control system.) subversion.apache.org. [Accessed January 30, 2014]

Bizon C, Ahalt S, Fecho K, Nassar N, Schmitt CP, Scott E, Wilhelmsen KC. Technologies for Genomic Medicine: CANVAS and AnnoBot, Solutions for Genomic Variant Annotation. RENCI Technical Report Series, TR-14-04. University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. doi: 10.7921/G0QN64N3, www.renci.org/TR-14-04. [Accessed March 27, 2014]

CANVAS (CAroliNa Variant Annotation Store), www.renci.org/TR-14-04 [Accessed March 27, 2014]

CASAVA (Consensus Assessment of Sequence and Variation). (A software program that converts raw sequencing data into a variety of formats used for downstream analysis.) www.illumina.com/software/genome_analyzer_software.ilmn. [Accessed January 6, 2014]

Chrome development tools. (Tools for Web development using the Chrome browser.) www.google.com/intl/en/chrome/browser. [Accessed January 30, 2014]

CLIA (Clinical Laboratory Improvements Amendments). www.fda.gov/medicaldevices/deviceregulationandguidance/ivdregulatoryassistance/ucm124105.htm. [Accessed December 13, 2013]

ClinGen (Clinical Genome Resource). (An NIH-funded resource that is under development to design and implement a framework for reaching expert consensus on delineating genomic variants that play a role in disease and those that are relevant to patient care.) www.iccg.org/about-the-iccg/clingen. [Accessed January 6, 2014]

ClinVar (Clinical Variants Resource). (A public archive of annotated genomic variants and associated clinically relevant effects on phenotype.) www.ncbi.nlm.nih.gov/clinvar/intro. [Accessed January 6, 2014]

Daemon. (A software program that runs as a background application and is disconnected from the user.) en.wikipedia.org/wiki/Daemon_%28computing%29. [Accessed January 6, 2014]

Data and Informatics Working Group, National Institutes of Health BD2K Initiative. NIH Request for Information: Management, integration, and analysis of large biomedical datasets. Analysis of public comments, 2012. NOT-OD-12-032. acd.od.nih.gov/DIWG_RFI_FinalReport.pdf. [Accessed October 31, 2013]

dbSNP (Single Nucleotide Polymorphism Database). (A public database containing species-specific, non-redundant sequence variations [i.e., SNPs, insertions, deletions, and short tandem repeats], as well as genotypes derived from the international HapMap project. Developed and maintained by the National Center for Biotechnology Information.) www.ncbi.nlm.nih.gov/SNP. [Accessed December 13, 2013]

Eclipse IDE (Integrated Development Environment). (A multi-language IDE that is comprised of a base workspace and extensible plug-in environment.) www.eclipse.org. [Accessed January 6, 2014]

ELSI (Ethical, Legal, and Social Implications) Research Program. (An NIH-funded program that was established in 1990 as part of the Human Genome Project to foster research on the ethical, legal, and social implications of genomics research.) www.genome.gov/elsi. [Accessed January 6, 2014]

Epic Systems Corporation. (A commercial EMR and Medical Billing System.) www.epic.com. [Accessed January 6, 2014]

ESP (Exome Sequencing Project). (A project that aims to establish a public database for genes that contribute to heart, lung, and blood disorders using data derived from richly-phenotyped patient populations. Funded by the National Heart, Lung, and Blood Institute, with contributions from numerous academic institutions.) evs.gs.washington.edu/EVS. [Accessed January 30, 2014]

Evans J, Berg JS. Genomic incidental findings: metaphors and methods. GENEWATCH 2014. www.councilforresponsiblegenetics.org/genewatch/GeneWatchPage.aspx?pageId=428. [Accessed January 7, 2014]

FastQC. (A software program that can be used to generate fastq files, which were developed by the Sanger Institute for efficient coding of sequencing data.) www.google.com/search?q=fastqc&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a. [Accessed February 10, 2014]

Firefox FireBug. (A Web development toolkit designed for integration with the Firefox browser.) getfirebug.com. [Accessed January 30, 2014]

Foreman AK, Lee K, Evans JP. The NCGENES project: exploring the new world of genome sequencing. NC Med J. 2013;74(6):500–504. www.ncmedicaljournal.com/archives/?74610. [Accessed February 6, 2014]

Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data. Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data. A white paper. Global Alliance; 2013. www.broadinstitute.org/files/news/pdfs/GAWhitePaperJune3.pdf. [Accessed June 16, 2013]

Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O'Daniel JM, Ormond KE, Rehm HL, Watson MS, Williams MS, Biesecker LG. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013;15(7):565–574. www.ncbi.nlm.nih.gov/pmc/articles/PMC3727274/pdf/nihms488659.pdf. [Accessed February 6, 2013]

HGNC (HUGO Gene Nomenclature Committee) database. (An NHGRI-funded committee that has been charged with approving unique symbols and names for human genes.) www.genenames.org. [Accessed January 30, 2014]

HGMD® (Human Gene Mutation Database). (A public database that contains published data on genomic variants that are associated with human inherited disease. Developed and maintained by the Institute of Medical Genetics in Cardiff University.) www.hgmd.cf.ac.uk/ac/index.php. [Accessed January 30, 2014]

Horvitz E, Mitchell T. From data to knowledge to action: a global enabler for the 21st century. Computing Community Consortium, v. 11. September 11, 2010. www.cra.org/ccc/docs/init/From_Data_to_Knowledge_to_Action.pdf.[Accessed March 11, 2013]

iRODS (integrated Rule-Oriented Data System). (An open source, policy-based data grid to integrate and manage data across distributed systems. Developed by the Data Intensive Cyber Environments group at UNC and the University of California at San Diego, with contributions from RENCI and other groups throughout the world. Maintained and distributed by the iRODS Consortium.) [www.irods.org/. [Accessed January 7, 2014]

JQuery. (A JavaScript library that is fast, small, and rich in features.) jquery.com. [Accesssed January 30, 2014]

JQWidgets (JQuery widgets). (A complete framework for building professional Web sites and mobile applications that is built upon open source technologies such as jQuery, Javascript, HTML5, and CSS.) www.jqwidgets.com. [Accessed January 30, 2014]

Kahn SD. On the future of genomic data. Science. 2011;331(6018):728–728.

Koboldt DC, Ding L, Mardis ER, Wilson RK. Challenges of sequencing human genomes. Brief Bioinf. 2010;11(5):484–498.

Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9:387–402.

Microsoft IIS (Internet Information Services). (An open architecture, flexible, secure Web server.) www.iis.net. [Accessed January 30, 2014]

Microsoft SQL (Structured Query Language) Server. (A cloud-based database server.) www.microsoft.com/en-us/sqlserver/product-info.aspx. [Accessed January 30, 2014].

Moore RW, Marciano R. Prototype preservation environments. LIBRARY TRENDS. 2005;54(1):144–162.

MySQL (Structured Query Language) (A variation of the Microsoft SQL Server.) www.mysql.com. [Accessed January 30, 2014]

NIH Staff. New NIH-funded resource focuses on use of genomic variants in medical care. September 25, 2013. www.nih.gov/news/health/sep2013/nhgri-25.htm. [Accessed January 6, 2014]

Oracle. (A cloud-based database server.) www.oracle.com/index.html. [Accessed January 30, 2014]

PHP (Hypertext Preprocessor). (An open source, general-purpose software language that is commonly used for Web development.) www.php.net/manual/en/intro-whatis.php. [Accessed January 30, 2014]

PostgreSQL pgAdmin. (An open source administration and development platform for managing the PostgreSQL database.) www.pgadmin.org. [Accessed January 30, 2014]

PostgreSQL. (An open source, enterprise-class, object-relational, SQL database system.) www.postgresql.org. [Accessed January 6, 2014]

python™. (An open source programming language designed to enable system integration. Developed and distributed by the Python community of software developers.) www.python.org. [Accessed December 13, 2013]

Rajasekar A, Moore R, Hou CY, Lee CA, Marciano R, De Torcy A, Wan M, Schroeder W, Chen SY, Gilbert L, Tooby P, Zhu B. iRODS Primer: integrated Rule-Oriented Data System. Synthesis Lectures on Information Concepts, Retrieval, and Services. 2010a;2 (1):1–-143.

Rajasekar A, Moore R, Wan M, Schroeder W, Hasan A. Applying rules as policies for large-scale data sharing. Proceedings of the UKSim/AM SS First International Conference on Intelligent Systems, Modelling and Simulation (ISMS), 2010b, pp. 322-327. Washington, DC, USA: IEEE Computer Society Washington.

REDCap™. (A Web-based software application to securely manage databases using audit trails and automated reports. Developed by Vanderbilt University and distributed by the REDCap™ Consortium.) www.project-redcap.org. [Accessed January 6, 2014]

RefSeq (Reference Sequence Collection). (A public database containing annotated, species-specific, non-redundant sequences [i.e., SNPs, insertions, deletions, and short tandem repeats], including genomic DNA, RNA, and proteins. Developed and maintained by the National Center for Biotechnology Information.) www.ncbi.nlm.nih.gov/refseq. [Accessed December 13, 2013]

Reilly J, Ahalt S, Fecho K, Jones C, McGee J, Roach J, Schmitt CP, Wilhelmsen K. Technologies for Genomic Medicine: MaPSeq, A Computational and Analytical Workflow Manager for Downstream Genomic Sequencing. RENCI Technical Report Series, TR-14-03. University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. doi: 10.7921/G0VD6WCF. www.renci.org/TR-14-03. [Accessed March 27, 2014]

Schmitt C, Wilhelmsen K, Krishnamurthy A, Ahalt S, Fecho K. Security and Privacy in the Era of Big Data: iRODS, a Technological Solution to the Challenge of Implementing Security and Privacy Policies and Procedures. RENCI/NCDS White Paper, Vol. 1, No. 2, 2013. University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. doi: 10.7921/GOH41PBR. www.renci.org/White-Paper-2013-iRODS. [Accessed November 27, 2013].

Shoenbill K, Fost N, Tachinardi U, Mendonca EA. Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. J Am Med Inform Assoc. 2014;21(1):171–180.

Sparx Enterprise Architect. (A modeling and design toolset.) www.sparxsystems.com. [Accessed January 30, 2014]

TeraGrid™ (A community of integrated, high-performance computers, data resources, and tools). http://info.teragrid.org/

vcf (variant call format). (A standard file format for storing genomic variant data and metadata.) www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41. [Accessed February 10, 2014]