Technologies for Genomic Medicine

CANVAS and AnnoBot, Solutions for Genomic Variant Annotation

A RENCI TECHNICAL REPORT
TR-14-04

CONTACT INFORMATION
Christopher Bizon, PhD
Telephone: 919.445.9600
Email: bizon@renci.org

List of Technical Terms and Websites

1000 Genomes Project, www.1000genomes.org

AnnoBot (Annotation Bot)

BioPython software, biopython.org/wiki/Main_Page

BLAT (BLAST-like Alignment Tool), www.blat.net

CANVAS (CAroliNa Variant Annotation Store)

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

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

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

gbff (GenBank flat file) format, www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html.

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

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

PostgreSQL (Structured Query Language) database, www.postgresql.org

python™ modules, www.python.org

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

SQLite3 database, www.sqlite.org

Introduction

Genomic medicine holds great promise to transform the medical profession and individualize health care. Technological advancements such as massively parallel genomic sequencing have made it possible to produce large amounts of genomic data within a reasonable timeframe and at a relatively low cost (Mardis, 2008; Horvitz and Mitchell, 2010; Koboldt et al., 2010; Kahn, 2011).

Projects such as the ClinVar and ClinGen initiatives, funded by the National Institutes of Health (NIH), are expanding our understanding of the clinical significance of genomic data through the adjudication of genomic variants and the methodical annotation of the genome (NIH Staff, 2013). Yet challenges remain in how best to interpret, reuse, and share the data (Ahalt et al., 2014; Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data, 2013; Data and Informatics Working Group, NIH BD2K Initiative, 2012).

Those challenges include the need for new technologies to capture, store, and update annotations to provide critical clinical interpretations of genomic data and metadata to attribute provenance or "ownership" and the history of a given data set (e.g., biological sources, laboratory processing steps, transformation and analysis steps, estimates of validity and reliability, etc.).

Herein, we describe two solutions—CAroliNa Variant Annotation Store (CANVAS) and Annotation Bot (AnnoBot)—that together provide version-controlled annotation and metadata to aid in the clinical interpretation of genomic variant data.

The Team*

Christopher Bizon, PhD, RENCI Senior Informatics Scientist; **Stanley Ahalt**, PhD, RENCI Director and Professor, Department of Computer Science UNC Chapel Hill; **Karamarie Fecho**, PhD, Medical and Scientific Writer for RENCI; **Nassib Nassar**, RENCI Senior Research Scientist; **Charles P. Schmitt**, PhD, RENCI Chief Technical Officer and Director of Informatics and Data Science; **Erik Scott**, RENCI Senior Research 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-14-04 in the RENCI Technical Report Series, May 2014
Renaissance Computing Institute
University of North Carolina
at Chapel Hill
919-445-9640
www.renci.org
@renci

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.

*Christopher Bizon serves as the technical lead on CANVAS and AnnoBot; Kirk Wilhelmsen serves as Principle Investigator and Director of RENCI's Biomedical Research division, which is leading the development of CANVAS and AnnoBot; all other team members are listed alphabetically

CANVAS[1]

CANVAS was developed initially to support a National Institutes of Health–funded research project, entitled “North Carolina Clinical Genomic Evaluation by NextGen Exome Sequencing” (NCGENES; Foreman et al., 2013). NCGENES is based at the University of North Carolina at Chapel Hill (UNC) and aims to explore the use of whole exome sequencing data for genomic research and clinical care. In order to achieve the study aim, RENCI needed to develop methods and approaches to: (1) match genomic variant data derived from NCGENES with reference genome data derived from publicly available databases; and (2) store the variant data with complete, up-to-date, version-controlled annotation derived from the reference genome data and other sources of variant annotation.

CANVAS was developed as a solution to these challenges. It is an open source, relational PostgreSQL database that stores genomic variant data with its associated annotation and metadata. CANVAS is designed as a relational data representation system that supports the management, query, and analysis of gigabyte- to terabyte-sized data sets from patient-level genomic sequencing data. CANVAS consists of ~85 tables organized into various schemas (Figure 1), including var (research project–specific variant data), refseq (reference genome data derived from RefSeq), and a collection of schemas for capturing additional variant annotation. Those annotation schemas are derived from several updateable data sources: dbSNP; the 1000 Genomes Project; ESP; HGNC; HGMD®; and ClinVar.

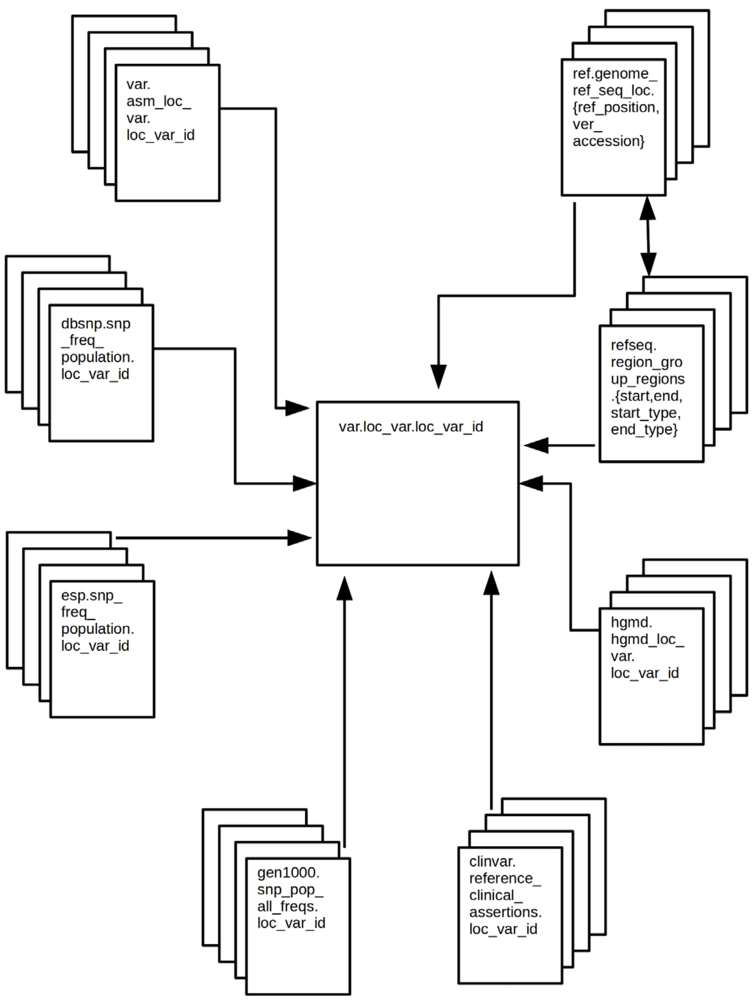

Figure 1. A schematic showing a high-level overview of the CANVAS architecture. Shown are schemas var, refseq, dbsnp, esp, gen1000, hgnc, hgmd, and clinvar. Not shown is the ClinBin schema.

Note that not only are there multiple annotation sources, with annotations organized and presented differently across sources, but the annotation sources are frequently updated as new information becomes available; this presents a challenge in how to ensure that the annotation in CANVAS is current and that all previous versions of annotation remain accessible in order to guide interpretation of current and past findings. AnnoBot (described below) was developed to monitor these external data sources for updates, extract any new annotation, and add that annotation to CANVAS. AnnoBot adds versioning information to all annotation to ensure that interpretations of genomic variant data are based on known data sources.

[1] CANVAS was formerly termed VarDB (Variant DataBase).

CANVAS also contains a schema ClinBin (not shown in Figure 1), which is used for NCGENES-specific computation to determine whether variants should be sorted into the diagnostic bin (DxBin) or incidental bin (IncidentalBin). DxBin includes variants that were targeted for a given patient/subject on the basis of a defined phenotype, have established clinical validity and utility, and thus are used for clinical diagnosis; in contrast, IncidentalBin includes incidental findings, or variants that were identified as a result of the sequencing effort but are believed to be unrelated to the disease phenotype or diagnostic goals and are thus used for research purposes only (Shoenbill et al., 2014, Foreman et al., 2013). ClinBin contains definitions of the variants in each bin, with annotation on the sources of those definitions, as well as data on patient-specific variants and their clinical significance. Of note, the variant data that are pushed into DxBin and IncidentalBin contain metadata on the origin of the variant annotation (i.e., they are version-controlled).

Storing annotation in CANVAS

Variant annotation is stored in Table var.loc_var, which contains loc_vars or variants located with respect to a specific reference sequence and described as follows:

- loc_var_id: an arbitrary integer identifier (surrogate key) assigned to the variant by the database
- pos: the position of the variant on the chromosome
- ref_id: the identifier of the reference sequence
- ref_ver_accession: the chromosome accession number
- ref: the reference allele
- alt: the alternate allele
- type: the variant type (i.e., single nucleotide polymorphism [SNP], insertion, deletion, substitution).

Acquiring output on genomic variants

An example of some of the query output provided by CANVAS for annotation on a given variant is shown below. Note that CANVAS provides approximately 70 fields of data on each variant, including data on the reliability and validity estimates for both the project's sequencing results and any reference data derived from public sources (see Owen et al., 2014 for additional fields and visual displays of output).

- **Variant id**: 57483
- **Chromosome:** 11
- **Position:** 1308393-1308394
- **Reference:** NCBI build 37.1
- **Analysis type:** incidental
- **Variant Class:** SNP
- **Variant:** A/T
- **Strand:** reverse
- **Minor Allele Frequency:** unknown
- **Zygosity:** heterozygous
- **Protein-coding effect:** missense
- **Gene:** thromboxane A2 receptor
- **Phenotype:** unknown
- **dbSNP ID:** rs5743
- **RefSeq ID:** NG_013363.1

Defining a variant's chromosomal position

CANVAS defines a variant's position on a chromosome as the physical location of a base in the reference sequence that is affected by the variant. A variant position p is the position between the physical locations p – 1 and p, for all p > 1; variant position p = 1 is the position preceding physical location 1. The bases displaced by a variant are those immediately following the variant position.

Reconciling ambiguous insertion/deletion variants

CANVAS invokes several functions to handle ambiguity in a variant's position. For example, consider a reported deletion of CAG from the beginning of the reference sequence CAGCAGCAG, which produces CAGCAG. Note that a deletion of AGC from the second position of the reference sequence, or a deletion of GCA from the third position, also produces CAGCAG. Because of this ambiguity, CANVAS describes the variant in a general or canonical form as a deletion replacing CAGCAGCAG with CAGCAG. This function is described as var.generalize_variant(). When implemented, the reference sequence is scanned to the left and the right of the ostensible variant location to determine if alternative candidate variants could produce the same sequence (i.e., ambiguity). If alternative candidate variants exist, then a single insertion/deletion is produced in the canonical form.

Adding variants to CANVAS

The database function var.loc_var_register() is used to add a variant to CANVAS. If the variant is an insertion or deletion, and if there could be ambiguity about its location due to sequence repeats, then this function expands the variant to a longer canonical form through the function var.generalize_variant(). If the variant already exists in the database, then the database returns its loc_var_id; otherwise, the variant is added to CANVAS, and the database generates a new loc_var_id.

AnnoBot

AnnoBot is a set of python™ modules and software driver code that are designed to automatically monitor targeted databases for updated information, extract new or revised annotations, and add those annotations to CANVAS. As noted above, the data sources that are currently monitored are: dbSNP, 1000 Genomes, ESP; HGNC; HGMD®; ClinVar; and RefSeq (Figure 1). AnnoBot can be extended to monitor additional databases as they become available.

AnnoBot implements the following python™ modules:

- Downloader: identifies and downloads new or edited annotations from external database sources
- Processor: transforms the data using BioPython
- Dbloader: uploads the data into CANVAS
- Mapper: maps the data to the genome using BLAT
- Maploader: filters the mapping before uploading it into CANVAS.

Describing AnnoBot's functionality using RefSeq as the primary external data source

RefSeq contains genomic sequencing data derived from many different species and stores this information as gbff files. The RefSeq Downloader reads each gbff file in this directory and identifies the ones derived from human data using the regular expression ORGANISM*Homo sapiens. RefSeq version numbers are indicated in the file names, and the Downloader captures this information as well.

The gbff files in RefSeq are hierarchical and similar to xml files. The RefSeq Processor invokes BioPython software to process the data files. BioPython parses the gbff files and transforms them into database-appropriate formats.

The RefSeq Dbloader implements a process similar to OR (Object Relational) mapping (Ambler, undated) to upload the transformed data into CANVAS. The Dbloader executes specific tasks, such as autoincrement counting, checking for existing rows, and maintaining links between the different tables in CANVAS.

The RefSeq Mapper uses BLAT to map the variant transcripts to the reference genome. BLAT conducts a gapped alignment; gaps in the alignment correspond to introns in the variant sequences. The RefSeq Maploader then uploads the mappings into CANVAS.

The AnnoBot Driver is used to manage the python™ modules and is comprised of software driver code and a SQLite3 database. The SQLite3 database organizes the state of each module, using hard versioning. The driver code continuously loops over the modules, processes any unprocessed data, uploads new processed data to the SQLite3 database, and invokes the Mapper to upload the new data to CANVAS.

A key feature of CANVAS and AnnoBot is annotation version control. As AnnoBot pulls updated annotation from external databases into CANVAS, the new annotation is stored in parallel with the older versions within the CANVAS schema. AnnoBot also pulls the version of the source data from which the annotation was derived. Version control allows the user to compare results across data and annotation sources—a feature that is missing from most other annotation systems.

Conclusion

CANVAS and AnnoBot work synergistically to provide a comprehensive solution to the challenges involved in maintaining a detailed, up-to-date, version-controlled record of genomic variant annotation, including metadata to record provenance and the history of a given data set.

Key Features:

- Architecture is open source
- Annotations are updated automatically
- All annotation is versioned and stored in parallel
- Queries are rapid and return rich output
- The system is modifiable and extendable
- The approach is generalizable to non-genomic annotation systems

Underlying Software and Technologies:

- BioPython
- BLAT
- PostgrSQL database
- python™
- SQLite3 database

Impact

- Currently supports variant annotation for the following research programs: (1) 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; (2) 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; (3) 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; and (4) UNCSeq, which applies tumor sequencing technology for >2,000 patient samples in order to identify mutations that are amenable to targeted treatments.
- Also supports the NIH-funded ClinGen and ClinVar initiatives (Dr. Jonathan Berg, Site PI), which involve a national effort to develop

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.

Karen Green provided editorial and design support for the preparation of this technical report.

How to reference this technical report:

Bizon, C., Ahalt, S., Fecho, K, Nassar, N., Schmitt, C., Scott, E., and Wilhelmsen, K. (2014): Technologies for Genomic Medicine: CANVAS and AnnoBot, Solutions for Genomic Variant Annotation. RENCI, University of North Carolina at Chapel Hill. Text. http://dx.doi.org/10.7921/G0QN64N3

renci

RESEARCH ∖ ENGAGEMENT ∖ INNOVATION

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]

Ambler S. Mapping objects to relational databases: O/R mapping in detail. www.agiledata.org/essays/mappingObjects.html. [Accessed December 16, 2013]

BioPython. (A set of open source software tools for biological computation. Developed and distributed by Python, an international team of software developers.) biopython.org/wiki/Main_Page. [Accessed December 13, 2013]

BLAT (BLAST-like Alignment Tool). www.blat.net. (An open source sequence alignment tool that can be used to identify the chromosomal position of a genomic sequence. Developed and distributed by the Genome Bioinformatics program at the University of California at Santa Cruz.) [Accessed December 13, 2013]

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 Polymorphisms 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]

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 that National Heart, Lung, and Blood Institute, with contributions from numerous academic institutions.) evs.gs.washington.edu/EVS. [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]

gbff (GenBank flat file) format. (A file format used by GenBank, which is an annotated collection of publicly available DNA sequences. Developed and maintained by the National Center for Biotechnology Information). www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html [Accessed May 15, 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]

HGNC (HUGO Gene Nomenclature Committee) database. (An National Human Genome Resource Institute–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 at 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]

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

Owen P, Ahalt S, Berg J, Coyle J, Evans J, Fecho K, Gillis D, Schmitt CP, Young D, Wilhelmsen. Technologies for Genomic Medicine: The GMW, a Genetic Medical Workflow Engine. RENCI Technical Report Series, TR-14-02. University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. doi: 10.7921/GOKW5CXC. www.renci.org/TR-14-02. [Accessed March 27, 2014]

PostgreSQL (Structured Query Language) database. (An open source, enterprise-class, object-relational 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]

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]

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;21:171–180.

SQLite (Structured Query Language Lite). (A set of open source software tools to implement a self-contained, serverless, zero-configuation, transactional SQL database engine. Developed and maintained by the SQLite Consortium.) www.sqlite.org. [Accessed December 13, 2013]

TeraGrid™. (A community of integrated, high-performance computers, data resources, and tools.) info.teragrid.org. [Accessed January 6, 2014]