BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3438–3439
doi:10.1093/bioinformatics/btu539

Databases and ontologies Advance Access publication August 14, 2014

Cordova: Web-based management of genetic variation data

Sean S. Ephraim[1,*], Nikhil Anand[1], Adam P. DeLuca[2], Kyle R. Taylor[3], Diana L. Kolbe[4], Allen C. Simpson[4], Hela Azaiez[4], Christina M. Sloan[4], A. Eliot Shearer[4,5], Andrea R. Hallier[1], Thomas L. Casavant[1], Todd E. Scheetz[1], Richard J. H. Smith[4,5,6,7] and Terry A. Braun[1,*]

[1]Department of Biomedical Engineering, [2]Department of Ophthalmology and Visual Sciences, [3]Department of Electrical and Computer Engineering, [4]Department of Otolaryngology—Head & Neck Surgery, Carver College of Medicine, [5]Department of Molecular Physiology & Biophysics, Carver College of Medicine, [6]Interdisciplinary Graduate Program in Genetics and [7]Iowa Institute for Human Genetics, Carver College of Medicine, The University of Iowa, Iowa City, IA 52242, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Cordova is an out-of-the-box solution for building and maintaining an online database of genetic variations integrated with pathogenicity prediction results from popular algorithms. Our primary motivation for developing this system is to aid researchers and clinician–scientists in determining the clinical significance of genetic variations. To achieve this goal, Cordova provides an interface to review and manually or computationally curate genetic variation data as well as share it for clinical diagnostics and the advancement of research.

Availability and implementation: Cordova is open source under the MIT license and is freely available for download at https://github.com/clcg/cordova.

Contact: sean.ephraim@gmail.com or terry-braun@uiowa.edu

Received on May 7, 2014; revised on June 22, 2014; accepted on August 4, 2014

1 INTRODUCTION

Curated genetic variation data play a crucial role for researchers, genetic counselors, clinicians and patients alike. We have created Cordova (curated online reference database of variation annotations), an out-of-the-box solution for building and maintaining a curated online database of genetic variations integrated with pathogenicity prediction results from popular algorithms. Cordova provides a collection of tools for research and clinical genetic testing designed to (i) annotate variations from popular pathogenicity prediction tools; (ii) collect published allele frequencies; (iii) provide an interface for reviewing and curating variation, pathogenicity and allele frequency data; and (iv) share these annotations to inform fellow clinicians and researchers with up-to-date data. We originally designed this software for dissemination of information on variations in genes causally related to non-syndromic hearing loss as part of the OtoSCOPE® diagnostic platform (Shearer *et al*., 2013). We then generalized the software to be applicable for management, curation and dissemination of any set of variations. The Leiden Open Variation Database (LOVD) (Fokkema *et al*., 2011) package is currently the only other software aimed at providing an out-of-the-box variation database management system. Therefore, we tested the latest version of the LOVD software (LOVD version 3, build 9) at the time of this writing and compared it with Cordova. A general comparison is provided in Table 1. LOVD is a popular and mature tool that stores variation data. However, we set out to develop a system that aids researchers and clinicians in the context of genetic testing and provides pathogenicity prediction results to aid in evaluation and determination of clinical significance of potentially disease-causing variants.

Cordova is written in PHP, built on the popular CodeIgniter 2 Web application framework and uses a MySQL database. Because Cordova is built with a popular Web framework, it is highly configurable and well-documented for development. LOVD is not based on a Web application framework. Our system provides a front-end Web interface for authenticated users to manage the database and a separate interface for browsing publicly available data. Standard Web security features are native to the CodeIgniter framework. In addition, Cordova includes a configurable Web API for quickly retrieving data in VCF, CSV, XML, JSON and tab-delimited formats. By comparison, the LOVD API output is not configurable and can only be viewed in Atom format. Current examples of Cordova installations include the Deafness Variation Database (DVD—http://deafnessvariationdatabase.org) and the Vision Variation Database (VVD—http://vvd.eng.uiowa.edu). The University of Iowa also hosts an internal Renal Disease Variation Database (RDVD) and a Dense Deposit Disease Database (DDDD) with plans for a public release in the future.

2 DATA MANAGEMENT

To submit a variation to a Cordova database, the user only needs to supply its genomic coordinates, reference allele(s) and alternate allele(s) (e.g. chr12:100751192:C>T). When a variation is first submitted, Cordova will attempt to auto-populate data from dbNSFP2 (Liu *et al*., 2013), Exome Variant Server (2013), The 1000 Genomes Project Consortium (2012) and OtoSCOPE®. Each allele frequency source can be configured for inclusion or exclusion. For example, the OtoSCOPE® source was excluded from the VVD. Auto-populated fields

*To whom correspondence should be addressed.

Table 1. Comparison of notable features and specifications for Cordova and LOVD

	Cordova	LOVD
Preliminary clinical significance	Yes	No
Data versioning	Yes	No
Web API output formats	VCF, CSV, XML, JSON, tab	Atom
Configurable Web API	Yes	No
Web framework	CodeIgniter 2	None
Version control	Git	SVN
Open source	GitHub	Trac
License	MIT	GPL

include gene symbol, variant locale, HGVS nucleotide change and amino acid change, functional prediction scores, conservation prediction scores and allele counts. Additional fields are supplied for clinical significance, phenotype, PubMed ID and general comments. Although any field can be edited, auto-populated fields are locked for editing by default to protect against submission of accidental misinformation. If a refSNP ID is supplied, a link to dbSNP will be provided automatically. Likewise, if a PubMed ID is supplied, a link to the corresponding PubMed entry will be included automatically.

For quality control, Cordova uses a queuing system, so users can choose to keep certain variations private while under review. When a user submits a new variation to the database, it will be entered into the queue and hidden from the public. If an existing variation is edited, any new changes will be held in the queue while older entries remain visible to the public. Variations in the queue can be released to the public at any time on review. Options are available to release either all variations in the queue at once or only those selected by the user. Users may also schedule a variation for removal from future database releases. Additionally, the database is robust to data versioning, so users can maintain copies of previous database releases and rollback to an earlier version if needed. Users who make queries via the Web API have the option to specify which version of the database to query. Data versioning is not supported by LOVD.

Cordova supports six pathogenicity prediction scores available from dbNSFP2, including SIFT (Kumar *et al.*, 2009), Polyphen-2 (Adzhubei *et al.*, 2010), MutationTaster (Schwarz *et al.*, 2010), LRT (Chun and Fay, 2009), phyloP (Siepel *et al.*, 2006) and GERP++ (Davydov *et al.*, 2010). Before manual curation, a variation can automatically be assigned a preliminary clinical significance classification based on the sum of these scores. If at least 60% of the available predictions have a pathogenic implication, a label of 'possibly pathogenic' will be assigned to that variation. Otherwise a label of 'predicted non-pathogenic' will be assigned. For variations with less than five prediction scores available, a label of 'unknown significance' will be assigned. Users may manually change the clinical significance label at any time. The prediction scores are also accompanied by a color-coded visual representation to easily understand the significance of each report. Red means the variant is predicted to be pathogenic, green means non-pathogenic and gray means that no score was provided from that particular algorithm. In comparison, LOVD does not offer a comparable feature for automated pathogenicity prediction, and therefore, users must manually provide clinical significance classifications.

Cordova comes packaged with an example pipeline to retrieve the necessary variation annotation data for auto-population. The pipeline is Ruby-based and can be used as a stand-alone command-line tool. Its output is configurable and can be tailored to local database needs, including customized interpretations of prediction and conservation scores from dbNSFP2. It uses compressed versions of annotation files to reduce their total footprint by one-third, and runs approximately eight times faster than dbNSFP2's native querying application for large sets of annotations.

Funding: This work was supported by the National Institutes of Health (NIH), National Institute on Deafness and Other Communication Disorders Grants DC011674 (A.E.S.), DC003544 and DC002842 (R.J.H.S.) and DC012049 (T.E.S., T.L.C., T.A.B., R.J.H.S.) and an NIH pre-doctoral research fellowship GM082729 (to A.P.D.).

Conflict of interest: none declared.

REFERENCES

Adzhubei,I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Chun,S. and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.*, **19**, 1553–1561.

Davydov,E.V. *et al.* (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.

Exome Variant Server. (2013) *NHLBI GO Exome Sequencing Project (ESP)*. Seattle, WA. http://evs.gs.washington.edu/EVS/ (7 January, 2014, date last accessed).

Fokkema,I.F. *et al.* (2011) LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.*, **32**, 557–563.

Kumar,P. *et al.* (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.

Liu,X. *et al.* (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.*, **34**, E2393–E2402.

Schwarz,J.M. *et al.* (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.

Shearer,A.E. *et al.* (2013) Advancing genetic testing for deafness with genomic technology. *J. Med. Genet.*, **50**, 627–634.

Siepel,A. *et al.* (2006) New methods for detecting lineage-specific selection. In: *Proceedings of the 10th Annual International Conference on Research in Computational Molecular Biology*. RECOMB, 2006, pp. 190–205. Springer-Verlag, Berlin, Heidelberg.

The 1000 Genomes Project Consortium et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.