Summer 2014

Design and application of methods for curating genetic variation databases

Sean Stephen Ephraim
University of Iowa

DESIGN AND APPLICATION OF METHODS FOR CURATING GENETIC VARIATION DATABASES

by

Sean Stephen Ephraim

A thesis submitted in partial fulfillment
of the requirements for the Master of
Science degree in Biomedical Engineering
in the Graduate College of
The University of Iowa

August 2014

Thesis Supervisor: Associate Professor Terry A. Braun

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

MASTER'S THESIS

_____

This is to certify that the Master's thesis of

Sean Stephen Ephraim

has been approved by the Examining Committee
for the thesis requirement for the Master of Science
degree in Biomedical Engineering at the August 2014 graduation.

Thesis Committee: _____
Terry A. Braun, Thesis Supervisor

_____
Todd E. Scheetz

_____
Richard J.H. Smith

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Terry Braun, for providing guidance, valuable feedback, overall direction, and, of course, companion trips to the coffee shop throughout my time spent in the Coordinated Laboratory for Computational Genomics. I would also like to thank Dr. Todd Scheetz and Dr. Richard Smith for providing a large wealth of assistance to help me complete this work. Thanks to Nikhil Anand and Dr. Eliot Shearer for establishing a foundation for me to build my work upon. Thanks to Dr. Thomas Casavant, Dr. Adam Deluca, Dr. Kyle Taylor, Dr. Diana Kolbe, Dr. Hela Azaiez, Allen Simpson, Christina Sloan, and Andrea Hallier for providing additional feedback and work to support my own work. Thanks to Bryce Diestelmeier and Sang Kyun Kang for aiding me in the process of writing this work. A very special thanks to my parents, Janice and Stephen Ephraim, for their endless support throughout the duration of my time in graduate school and, well, life. Additional thanks goes to all of my wonderful family and friends whose continuous encouragement has kept me motivated. Funding was provided by the National Institutes of Health (NIH), National Institute on Deafness and Other Communication Disorders Grants F30 DC011674 (AES), DC003544 (RJHS), DC002842 (RJHS) and DC012049 (TES, TLC, TAB, RJHS) and an NIH Pre-doctoral Research Fellowship T32 GM082729 (to APD).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Curated genetic variation data play a crucial role for researchers, genetic counselors, clinicians, and patients alike. This thesis describes 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: 1) annotate variations from popular pathogenicity prediction tools; 2) collect published allele frequencies; 3) provide an interface for reviewing and curating variation, pathogenicity and allele frequency data; and, 4) share these annotations to inform fellow clinicians and researchers with up-to-date data on variations. The initial implementation of this software was for dissemination of information on variations in genes causally related to only nonsyndromic hearing loss as part of the OtoSCOPE® diagnostic platform (Shearer *et al.*, 2013). For this thesis, the software was generalized 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 to Cordova. 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 provide reviewers of patient variation data with pathogenicity prediction results to aid with evaluation and determination of clinical significance of potentially disease-causing variants.

The Deafness Variation Database (DVD) is an online database of genetic variations in genes known to be associated with deafness. It was created by The Molecular Otolaryngology and Renal Research Laboratory at the University of Iowa. The

DVD's collection of genes was accumulated from the OtoSCOPE® study, and the variation annotations are based on publicly available data. The original version of the DVD was integrated with dbNSFP 1 and contained prediction scores from six algorithms. The release of dbNSFP 2 introduced structural changes to the data as well as minor alterations to the method of reporting prediction scores. Therefore, the DVD was ported to support the updated data for these same six prediction algorithms from dbNSFP 2. This thesis describes the design and implementation of DVD 3.0, which includes substantially more data for genes and variations as a result of the amount of genomic sequence information that is publically available. The DVD includes annotations from software-derived analyses and manual curation from human experts.

Chapter 2 describes the public resources used to acquire known disease-associated variants. Chapter 3 describes the methods behind the implementation of the DVD and a novel scoring method that integrates population-based minor allele frequencies, pathogenicity predictions and known disease alleles. Chapter 4 describes the Cordova system, and Chapter 5 summarizes the results.

CHAPTER 2: BACKGROUND

This chapter provides information on the tools, technologies, databases, and important concepts surrounding the basis of this work.

2.1 dbNSFP

dbNSFP (database for non-synonymous SNPs' functional predictions) is a collection of results from popular prediction algorithms (Liu *et al.*, 2011, 2013). dbNSFP 1 was introduced in 2011, and dbNSFP 2 was introduced in 2013. Since its release, dbNSFP 2 has seen several minor updates with a growing number of prediction algorithms. The most recent release at the time of this writing is dbNSFP 2.5 that contains prediction records for over eighty seven million nonsynonymous SNPs. dbNSFP contains algorithms for predicting whether a variation will be functional or deleterious as well as algorithms for predicting the conservation of a variation. Functional algorithms include SIFT (Kumar *et al.*, 2009), Polyphen-2 (Adzhubei *et al.*, 2010), LRT (Chun and Fay, 2009), MutationTaster (Schwarz *et al.*, 2010), MutationAssessor (Reva *et al.*, 2011), FATHMM (Shihab *et al.*, 2013), CADD (Kircher *et al.*, 2014), and VEST (Carter *et al.*, 2013). Conservation algorithms include phyloP (Siepel *et al.*, 2006), PhastCons (Siepel *et al.*, 2005), GERP++ (Davydov *et al.*, 2010), and SiPhy (Garber *et al.*, 2009).

dbNSFP is divided into twenty four text files where each file represents a single chromosome. Each line of a file contains the relevant information for a single nonsynonymous SNP and columns are tab-separated. dbNSFP 2 contains an additional file for gene information. The latest version of dbNSFP 2 is over 30GB in size. dbNSFP comes packaged with a native querying program written in Java that can be executed from the command line. This program contains basic options for specifying the input and output file names, specifying the preferred human genome reference sequence (e.g. hg19), specifying which chromosome files to search, and specifying the output columns.

In addition to reporting the raw prediction scores, dbNSFP 2.5 also reports normalized scores. Each algorithm produces a unique range of raw scores, but the normalized scores all scale from zero to one and serve as a means for comparison and interpretation. dbNSFP also contains nominal interpretations of the normalized prediction scores including classifications such as "Probably damaging", "Possibly damaging", and "Benign". dbNSFP 1 only reported the normalized scores and nominal interpretations and withheld the raw scores. dbNSFP 2.0 began reporting the raw prediction scores as well, but interestingly enough, the normalized scores and the nominal predictions for SIFT and phyloP were dropped. These two fields were added in release dbNSFP 2.1 for SIFT, but so far they have not been reintroduced for phyloP.

2.2 HGMD

The Human Gene Mutation Database (HGMD) is a collection of disease-causing and disease-associated genetic variations that was first introduced in 1996 (Stenson *et al.*, 2014). Users can pay a subscription fee to access quarterly releases of HGMD Professional, containing over 148,000 records of the most up-to-date data. A free option is also available to users in academics and non-profit organizations. However, this option is restricted to data from three years prior, has limited features, and is only released every six months. HGMD is commercially licensed through BIOBASE. BIOBASE additionally offers subscriptions to Genome Trax, a collection of several mutation databases including HGMD.

All data from HGMD are collected from published results in the scientific literature from over 1950 journals. Manual and automated screening methods are used to identify the integrity of a publication. In some cases, HGMD curators will contact the original authors for further clarification on the clinical significance of a mutation. HGMD has defined seven different variant classifications that are summarized in Table 1. Their

policy is to enter any variation into the database that has been associated with disease, even if the functional aptitude is unclear. All such variations are clearly indicated.

Table 1: Definitions for the classifications of variants in HGMD.

HGMD classification	**Definition**
DM	Disease-causing mutation
DM?	Probable/possibly disease-causing mutation; author expresses uncertainty
CNV	Copy number variations
FTV	Frameshift or truncating variant
FP	In vitro/laboratory or in vivo functional polymorphism
DFP	Disease-associated polymorphism with additional supporting functional evidence
DP	Disease-associated polymorphism

2.3 dbSNP and ClinVar

dbSNP is a publicly available genetic variation database that includes records of disease-causing and non-disease-causing variants (Sherry *et al.*, 2001) . Each SNP is given its own unique ID called a RefSNP ID. ClinVar is a sister variation database of dbSNP that contains additional information on clinical significance submitted by expert curators. While HGMD only contains variations that have been published in the scientific literature, ClinVar contains both published and unpublished variations (Stenson *et al.*, 2014). dbSNP and ClinVar are both hosted by the National Center for Biotechnology Information.

2.4 Locus-specific databases

A locus-specific database (LSDB) is a general term for a database typically containing variations associated with a gene, disease, set of genes or set of diseases. LSDBs are usually setup by individual curators. LSDBs are typically much smaller than HGMD, dbSNP, and ClinVar for this reason. While HGMD only contains variations that have been published in the scientific literature, LSDBs generally contain both published and unpublished variations (Stenson *et al.*, 2014).

2.5 The Leiden Open Variation Database

The Leiden University Medical Center has developed the Leiden Open Variation Database (LOVD) software to serve as what they call an "LSDB-in-a-box" (Fokkema *et al.*, 2011). LOVD is a freely available package that aims to allow curators to setup an online genetic variation database. The purpose of the software is to give curators an easily installable and user-friendly platform to collect, share, and curate genetic variants. Since the introduction of LOVD in 2005, many curators have adopted the system for their own LSDBs. The Leiden team has also setup many of their own LOVD installations and is currently seeking curators to manage them.

2.6 Minor allele frequency studies

OtoSCOPE®, 1000 Genomes, and Exome Variant Server (EVS) are all studies designed to gather data on minor allele frequencies (MAFs). The OtoSCOPE® dataset contains MAFs from six populations including Ashkenazi Jewish, Columbian, Japanese, European-American, Spanish, Turkish. It is an ongoing study that has revealed a total of eighty-three deafness-associated genes thus far. The 1000 Genomes dataset includes MAFs from four super-populations including African, American, Asian, and European. The EVS dataset was integrated for the retrieval of MAFs from two populations including European-American and African-American.

2.7 Technologies

PHP and Ruby are both popular software programming languages. Both of these languages have many uses, but PHP was selected to develop our web software and Ruby to develop our command line software. CodeIgniter 2 is a web application framework written in PHP used by developers to build robust websites and is also adopted for the implementation of Cordova. A web application framework is an infrastructure and set of web programming libraries that developers can use to organize their code. Additionally, HTML and CSS are both web languages as well. While PHP is responsible for handling the logic of websites, HTML and CSS are responsible for handling the presentation of websites. JavaScript is another web programming language, and in some cases it can handle website logic as well. However, JavaScript can only be used within a browser, and PHP can only be used within a server. For this reason, JavaScript is often referred to as a client-side language while PHP is often referred to as a server-side language. Additionally, MySQL is a language used to create and manage databases stored on servers.

The source code for software applications can be organized in many ways. There are many architectural patterns that developers can utilize to structure their code and functionality, but one popular choice is the model-view-controller (MVC) architectural pattern. As the name suggests, the code and functionality are split into three main components: the model, the view, and the controller. The view refers to the presentation layer that the user sees and interacts with. The model handles all interactions with the database including querying and updating records. The controller is the sole messenger between the view and the model. It directs all user interactions from the view to the model and back again. The CodeIgniter 2 web application framework implements the MVC architecture.

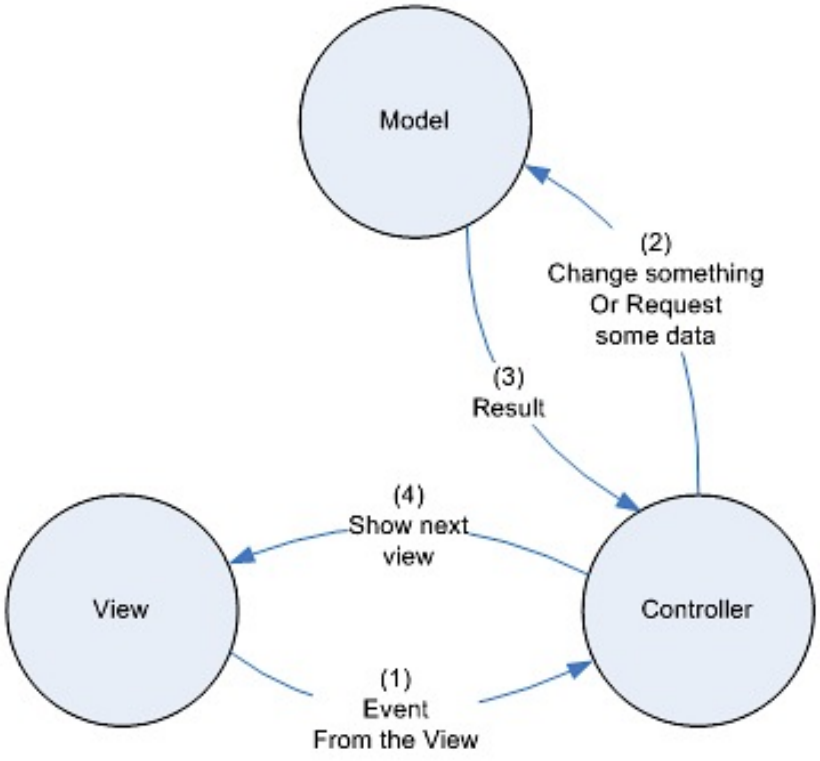

Figure 1: The model-view-controller (MVC) architectural pattern for the CodeIgniter 2 framework. It is also used in many software applications.

Source: http://blog.stannard.net.au/blog/media/simple-mvc-framework/mvc.gif

CHAPTER 3: THE DEAFNESS VARIATION DATABASE

This chapter provides an overview of the methods used to implement the DVD. The results are presented and followed by discussion.

3.1 Methods

3.1.1 Evaluation of variations from HGMD, dbSNP and ClinVar

Prior to implementing DVD 3.0, an evaluation of variations of the 69 genes in the OtoSCOPE® panel (July 2012) was performed. Data was acquired from HGMD, dbSNP and ClinVar. For this evaluation, all variations labeled "pathogenic" were obtained to compare the maximum MAF seen in any population to the sum of pathogenic predictions from six pathogenicity prediction algorithms as shown in Figure 3. The same comparison was made for variations labeled "non-pathogenic" as shown in Figure 4. Only variations with population data and scores from six pathogenicity prediction algorithms were used for this evaluation.

3.1.2 Collecting variation data for DVD 3.0

Data was collected from multiple databases to begin the process for assembling the DVD. Variant annotation data for genome build hg19 was gathered from dbSNP, ClinVar, HGMD, OtoSCOPE®, EVS, 1000 Genomes and dbNSFP 2. To do this, dbSNP, ClinVar and HGMD were first queried for each of the eighty-three deafness-associated genes identified in the OtoSCOPE® study. Genomic coordinates (hg19), reference alleles, alternate alleles, amino acid changes, clinical significance, refSNP IDs, PubMed IDs for these variants were collected from these databases, HGMD was accessed using the Genome Trax MySQL database. dbSNP was accessed by using the online query form at http://www.ncbi.nlm.nih.gov/SNP and downloading the results in XML format. ClinVar was accessed by downloading the database in VCF format via the FTP server at ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF. Next, MAFs for each of the

variants were collected from the OtoSCOPE®, EVS, and 1000 Genomes population databases. The OtoSCOPE® population data was acquired through the Molecular Otolaryngology and Renal Research Laboratory. The EVS data was accessed by downloading the database in VCF format from http://evs.gs.washington.edu/EVS. The 1000 Genomes data was accessed by downloading the database in VCF format via the FTP server at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp. Lastly, algorithmic prediction scores and interpretations were collected from dbNSFP 2. The dbNSFP 2 data was accessed by downloading the data in tab-delimited format from https://sites.google.com/site/jpopgen/dbNSFP. OtoSCOPE®, EVS, 1000 Genomes, and dbNSFP 2 databases were queried using the hg19 genomic coordinates, reference alleles, and alternate alleles from each of the variants were collected earlier. The full pipeline for collecting variant annotation data can be seen in Figure 2.

After collecting the annotation data for deafness-associated variants, the results were combined to form a single record for each variant. The phenotypic labels were mapped from HGMD and ClinVar to alternate labels standardized for the DVD. A full list of phenotypic label conversions can be seen in Table A-1. The clinical significance labels were standardized among HGMD, dbSNP and ClinVar by re-assigning “DM” and “DM?” labels in HGMD to “pathogenic” and “probable-pathogenic”, respectively.

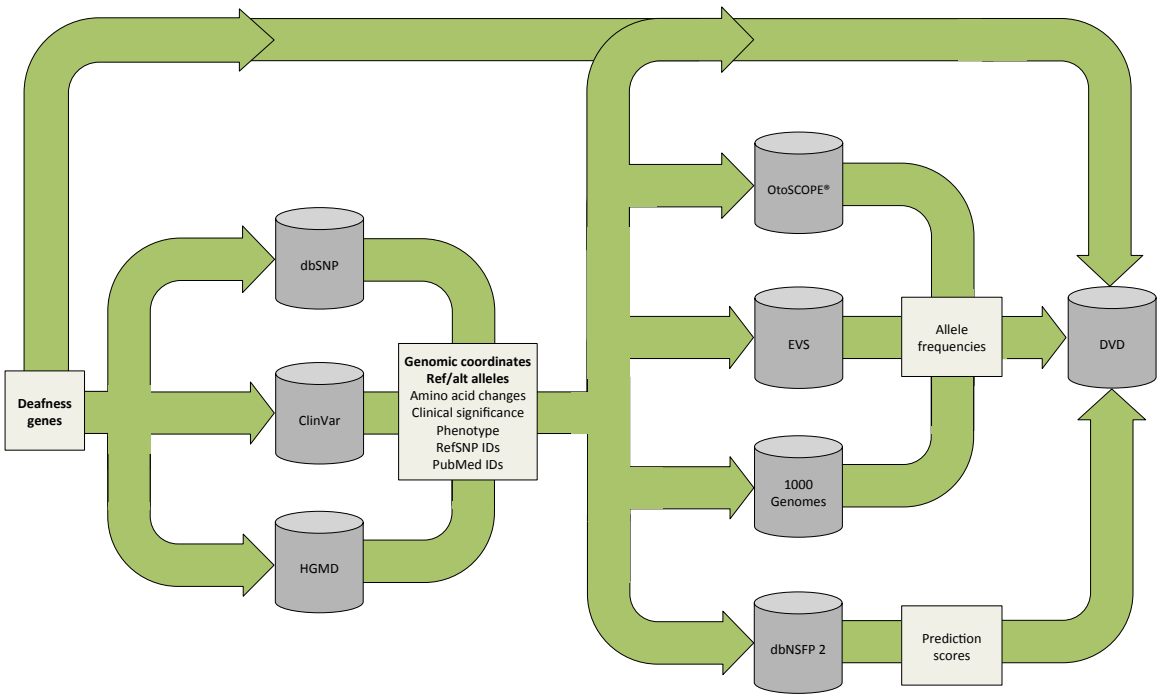

Figure 2: Annotation pipeline for accessing the dbSNP, ClinVar, HGMD, OtoSCOPE®, EVS, 1000 Genomes, dbNSFP 2 databases.

Note: variant attributes used to query each subsequent database are highlighted in bold text.

3.1.3 Ranking for non-pathogenicity

A novel scoring method was designed for ranking variants based on population-based MAFs, clinical significance, coding effects and pathogenicity predication and has been created. The PROVE (plausible re-classification of variant effect) index combines data from these different categories to generate an overall non-pathogenicity score. The approach of this index is to apply a score to each variant and then sort the scores from highest (least pathogenic) to lowest (most pathogenic) -- lower scores are more pathogenic. A variant's ranking increases when certain criteria are met, and variants ranked higher than others indicate a higher likelihood of being non-pathogenic. All

criteria can be seen in Table 2. The PROVE index is a conceptual result that remains to be implemented in software.

Table 2: Criteria for the assignment of a PROVE index to a variant.

Field	Criteria	Weight	Max weight
Population	MAF > 0.5% (AR), or MAF > 0.05% (AD)	100 per population	1200
Clinical significance from another database	dbSNP/ClinVar calls it benign	20	40
	HGMD calls it benign		
	dbSNP/ClinVar calls it likely benign	10	
	HGMD calls it likely benign		
Coding effect	Synonymous	8	8
	Missense	6	
	Stop loss/gain	4	
	Splice site	2	
Algorithmic prediction	Predicted benign/conserved	0.01 per algorithm	0.10

MAF, minor allele frequency; R, autosomal recessive; D, autosomal dominant.

Note: The index provides a weighted sum of the evidence for non-pathogenicity, but it does not provide detailed information about the origin of the evidence.

3.2 Results

3.2.1 Evaluation of variations from HGMD, dbSNP and ClinVar

The following results are from the evaluation of gene variations from HGMD, dbSNP and ClinVar for 69 OtoSCOPE® genes.

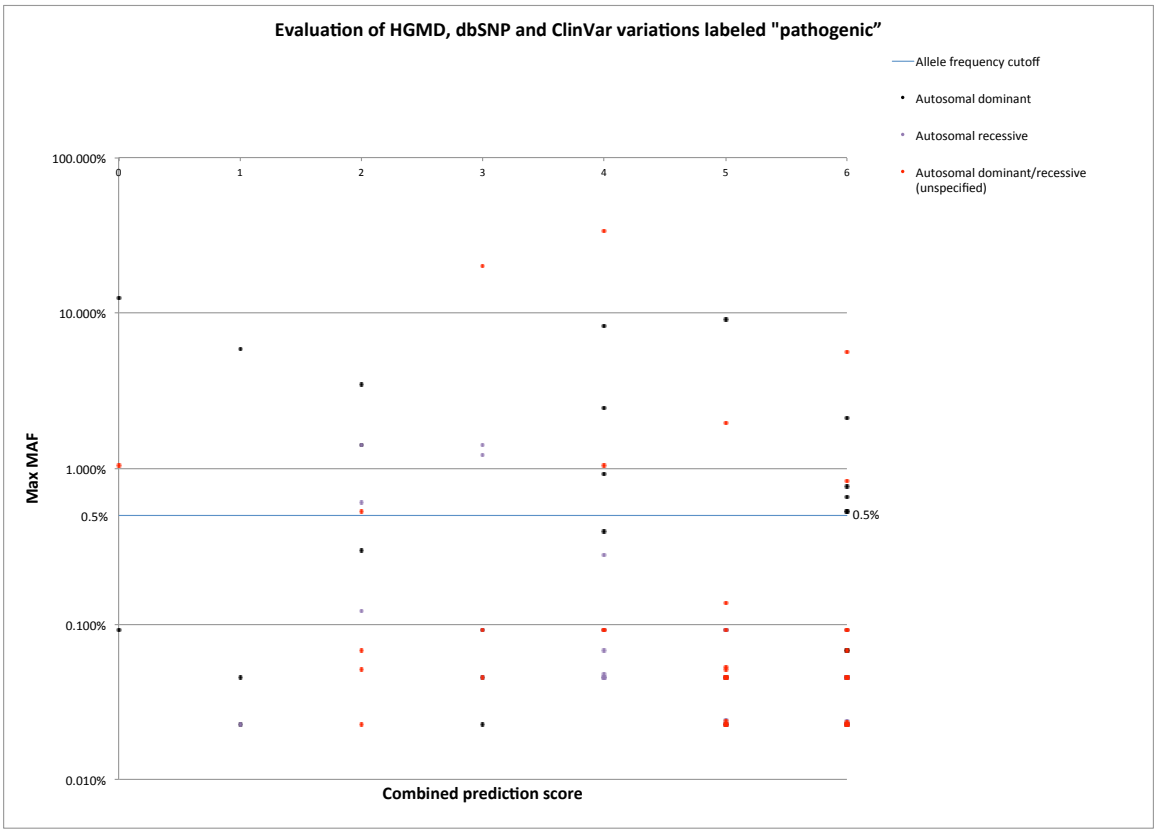

Figure 3: Evaluation of variations labeled "pathogenic" by HGMD, dbSNP and ClinVar. The maximum minor allele frequency observed in any of the twelve populations is mapped against the combined pathogenicity prediction score from 6 popular prediction algorithms.

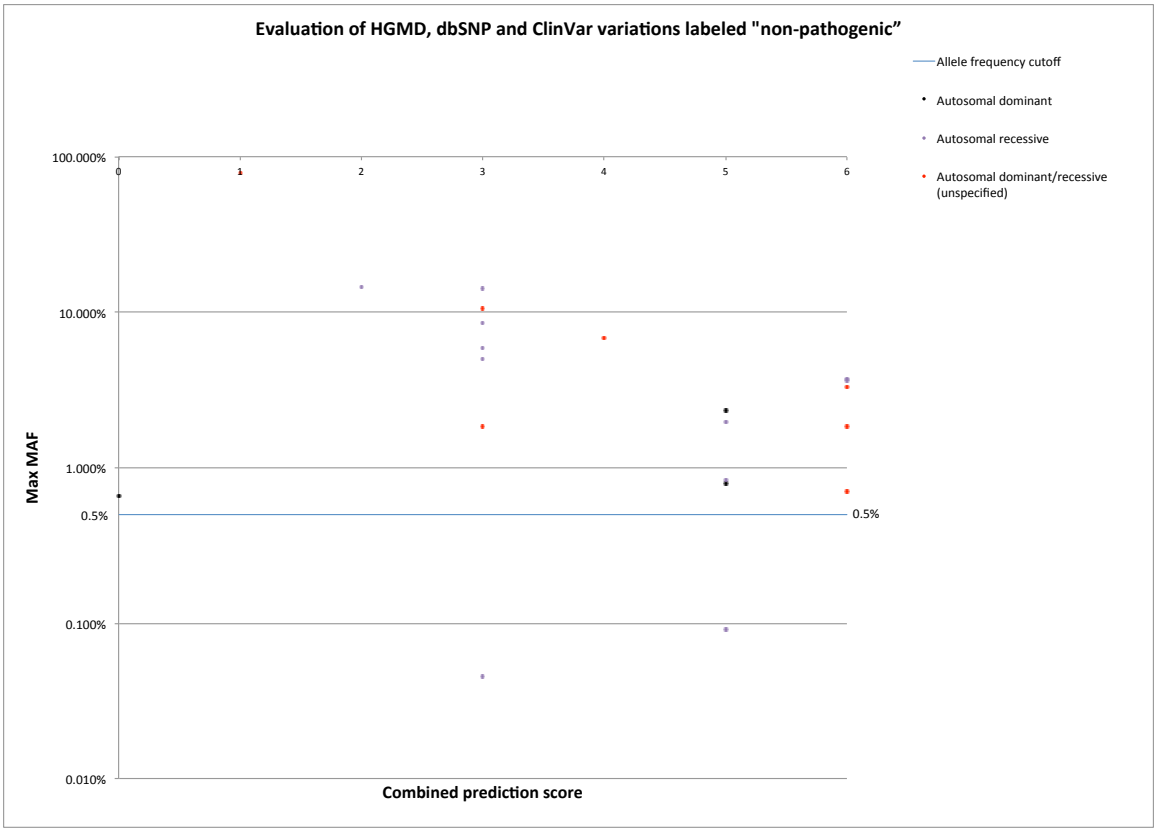

Figure 4: Evaluation of variations labeled "non-pathogenic" in HGMD, dbSNP and ClinVar. The maximum allele frequency of each variant is mapped against the combined pathogenicity prediction score from 6 popular prediction algorithms.

3.2.2 Collecting and analyzing variation data for DVD 3.0

A total of 137,579 records were collected from dbSNP, 2324 records from ClinVar, and 3260 records from HGMD for a total of 14,3163 records. After filtering for only variants labeled as pathogenic and combining records to produce one record per variant, there were a total of 3421 pathogenic variants. From these databases, inheritance pattern for 3109 variants could be identified. The literature was used to determine the inheritance pattern of the remaining 312 variants. A closer examination into the literature

for seventy GJB2 variants revealed four discrepancies between the stated clinical significance and the associated reference.

Table 3: Discrepancies between HGMD clinical significance and HGMD reference for GJB2 variants.

Nucleotide change	Amino acid change	HGMD classification	Discrepancy notes	HGMD reference
c.164C>A	p.T55N	DM	Publication states that this variant may "only be a polymorphism".	(Tekin *et al.*, 2005)
c.331A>G	p.I111V	DM	This was the only GJB2/DFNB1 mutation in this patient, meaning that the patient could be deaf due to other reasons.	(Azaiez *et al.*, 2004)
c.110T>C	p.V37A	DM		
c.61G>A	p.G21R	DM	No reference given. This conflicts with HGMD's protocol, which states that unpublished (non-peer reviewed) data will not be entered (Stenson *et al.*, 2014).	NA

DM, disease-causing mutation.

3.3 Discussion

The evaluation of data from HGMD, dbSNP and ClinVar for 69 OtoSCOPE® genes shown in Figure 3 and Figure 4 provided evidence to demonstrate that some variations are too common to be considered disease-causing. Our MAF populations included data from 1000 Genomes, EVS, and OtoSCOPE®. For this evaluation, based on the prevalence of hearing loss, a permissive MAF cutoff of 0.5% was chosen. Figure 4

shows that most known non-pathogenic variations had a MAF above this cutoff. Figure 3 shows that some pathogenically labeled variations not only had a MAF above this cutoff as well but were also predicted to be benign most or all of the time. We were particularly interested in such variations and collected up-to-date data to perform another evaluation. For the DVD 3.0, a MAF cutoff of 0.5% was chosen for autosomal recessive variations and a MAF cutoff 0.05% for autosomal dominant variations. HGMD performed a similar analysis to remove or re-categorize variant records. By comparison, they chose a cutoff of 1% for all variants and only used MAF data 1000 Genomes populations (Stenson *et al.*, 2014). This analysis identified a total of ninety-three pathogenically labeled variants as too common to cause disease, and were therefore re-classified as benign.

The PROVE index scores variants based on population data, inheritance patterns, algorithmic prediction scores, variant coding effects and previous clinical significance reports from HGMD, dbSNP, and ClinVar. This index aims to allow interpretation of five-dimensional data in a single dimension. A rank threshold for filtering can be chosen based on the user's discretion. Additionally, the rank number encodes information about the respective variant as seen in Figure 5.

In this study, there were up to ten prediction scores, twelve populations of MAFs, and three previously reported clinical significance values associated with each variant. The large volume of data can make it difficult to determine which variants need to be re-classified. The PROVE index was designed so that a larger index number would directly correlate to the amount of supporting evidence to re-classify from pathogenic to non-pathogenic. It is important to know that the PROVE index is merely a filtering strategy intended to flag variants for possible re-classification based on existing prediction, population, and clinical significance data. It is possible for one variant to have a higher PROVE index than another and still be pathogenic. A curator has the option to set a threshold value as high or as low as they want based on their discretion. While a MAF cutoff of 0.05% was used for autosomal dominant variants and 0.5% for autosomal

recessive variants, curators also have the option to choose their cutoff values. Altering the MAF cutoff values would ultimately alter each PROVE index as well.

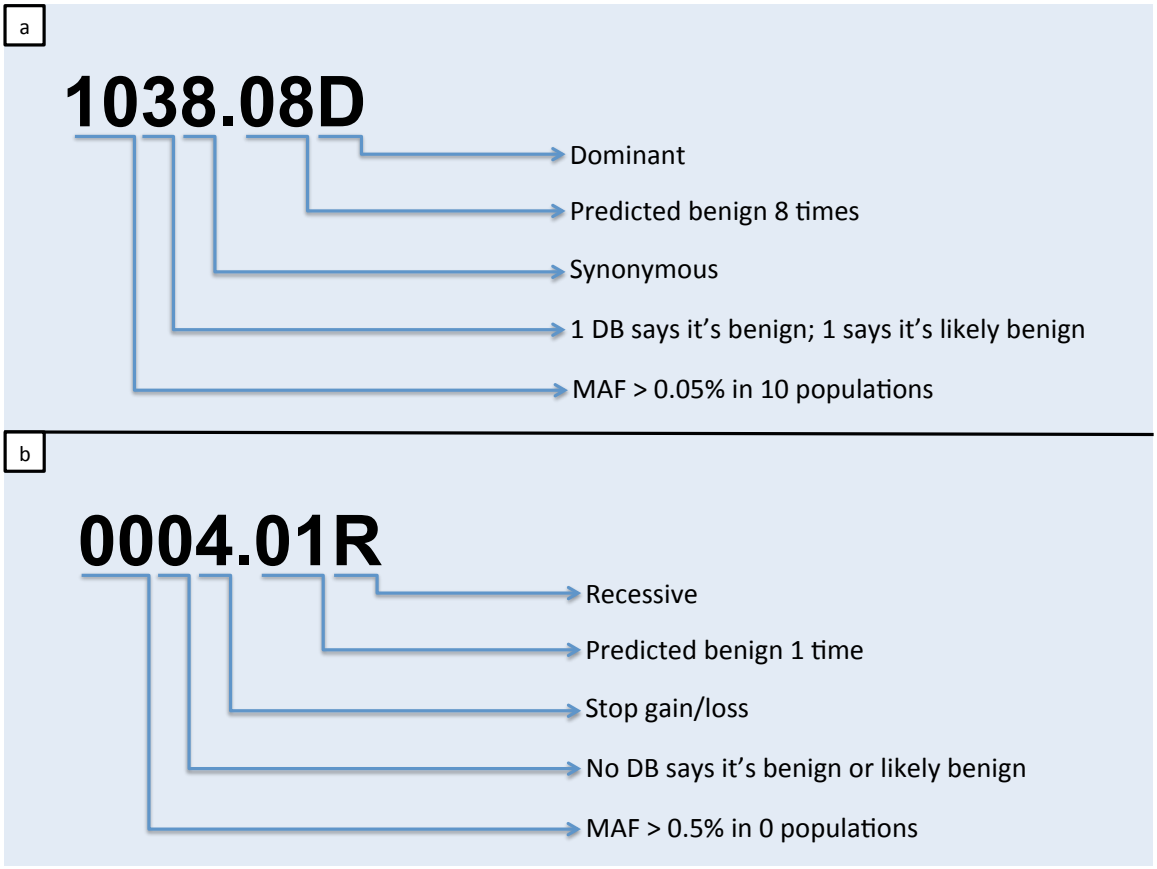

Figure 5: Examples of a variant's PROVE index and its interpretation. (a) A variant with a high index has more supporting evidence to suggest that it is benign than (b) a variant with a low index.

One problem that was encountered in acquiring data for the PROVE index was that the various databases were not synchronized with each other. For example, in the set of SNPs examined, over 300 records could be found through a dbSNP website query but the same records were not present in the VCF files of build 138 on the dbSNP FTP server. The same was true for the dbSNP table found on the UCSC website. For

example, rs267598119 could be found on the dbSNP website but could not be found within the VCF files even though records state that that variant was added in build 136 and updated in build 138. The missing records were confirmed by the administrators of dbSNP and UCSC and were corrected after this discrepancy was described to them.

CHAPTER 4: CORDOVA

This chapter provides an overview of the methods used to create the Cordova software. Cordova is an out-of-the-box solution for building, managing and maintaining a LSDB. It implements population data and scores from pathogenicity prediction algorithms to help researchers and clinicians determine the clinical significance of genetic variations. Screenshots from the software are presented and followed by discussion.

4.1 Methods

4.1.1 Web application

Cordova was built with several modern web development technologies. The PHP 5 web programming language was used to process server-side requests. Browser views were written in HTML and stylized with CSS. JavaScript and the JQuery 1.6 library were used to handle client-side (browser) requests. The Twitter Bootstrap 2 template library was also used to further enhance browser views and client-side requests. The entire Cordova infrastructure was built into the CodeIgniter 2 web application framework. Cordova was written specifically to be compatible with the MySQL 5 database management language. Most of the database queries were written using CodeIgniter's native database abstraction layer. However, due to complexity, some queries were custom written in MySQL. The Ion Auth 2 library was used for local authentication and the University of Iowa's Active Directory API was used for remote authentication. The dompdf 0.5 library was implemented to render a variation's results in PDF format. The pChart 2 library was used to render the bar charts for a variation's MAF data. A Google Analytics module was integrated for optional website traffic monitoring. Local version control was managed using Git and remote version control was hosted on GitHub. The official Cordova installation documentation was written in Markdown syntax. The source

code was documented using the syntax standards of phpDocumentor 2, and the class documentation was generated using phpDocumentor 2. Markdown and phpDocumentor 2 both produce standard HTML output.

The CodeIgniter 2 framework and the Twitter Bootstrap 2, JQuery 1.6, and Ion Auth 2 libraries were packaged directly with the Cordova repository. An auto-installation feature was written to fetch the dompdf 0.5 and pChart 2 libraries from their respective repositories and properly install them into Cordova without manual assistance.

4.1.2 Annotation pipeline

The configurable pipeline called Kafeen was built to complement Cordova by automatically retrieving relevant annotations. Kafeen consists of an algorithm and annotation datasets. Kafeen was developed using the Ruby programming language and integrated with existing third-party software including Trollop, ASAP, tabix (Li, 2011), and bgzip (Li, 2011). The Ruby library Trollop was incorporated to efficiently handle option-parsing from the command line. ASAP was integrated to allow Kafeen to retrieve the gene, HGVS protein change, HGVS nucleotide change and variant locale for each variant. tabix, a SAMtools (Li *et al.*, 2009) utility, was integrated to quickly retrieve variant annotations from large datasets. These datasets include dbNSFP 2, OtoSCOPE®, 1000 Genomes, and the ESP6500SI dataset from EVS. The dbNSFP 2 dataset was integrated for the retrieval of pathogenicity scores from prediction algorithms including SIFT (Kumar *et al.*, 2009), Polyphen-2 (Adzhubei *et al.*, 2010), LRT (Chun and Fay, 2009), MutationTaster (Schwarz *et al.*, 2010), phyloP (Siepel *et al.*, 2006), GERP++ (Davydov *et al.*, 2010). The OtoSCOPE® dataset was integrated for the retrieval of MAFs from six populations including Ashkenazi Jewish, Columbian, Japanese, European-American, Spanish, Turkish. The 1000 Genomes dataset was integrated for the retrieval of MAFs from four populations including African, American, Asian, and European. The EVS dataset was integrated for the retrieval of MAFs from two populations including

European-American and African-American. In order to prepare the datasets for use with tabix, the files are first converted to the Variant Call Format (VCF) (Danecek *et al.*, 2011). All files are then compressed using the bgzip compression tool packaged with tabix. To accommodate for the rather rapid release schedule of dbNSFP 2, the option was integrated to allow users to easily compile new versions of dbNSFP 2 for use with Kafeen.

Table 4: Criteria for converting dbNSFP 2 nominal and numerical prediction scores to a standard binary scale.

Algorithm		**dbNSFP 2 prediction value**	**Definition**	**Converted score**
Functional	LRT	D	Deleterious	1
		N	Neutral	0
		U	Unknown	None
	MutationTaster	D	Disease-causing	1
		A	Disease-causing (automatic)	
		N	Polymorphism	0
		P	Polymorphism (automatic)	
	Polyphen-2 (HDIV)	D	Deleterious	1
		P	Possibly damaging	
		B	Benign	0
	SIFT	< 0.05	Damaging	1
		≥ 0.05	Tolerated	0
Conservation	phyloP	> 1	Conserved	1
		≤ 1	Non-conserved	0
	GERP++	> 0	Conserved	1
		≤ 0	Non-conserved	0

4.2 Results

The following figures provide the results of the Cordova software. All screenshots were taken from the DVD website, which is a production instance of Cordova. Annotations are provided to highlight key features of the software.

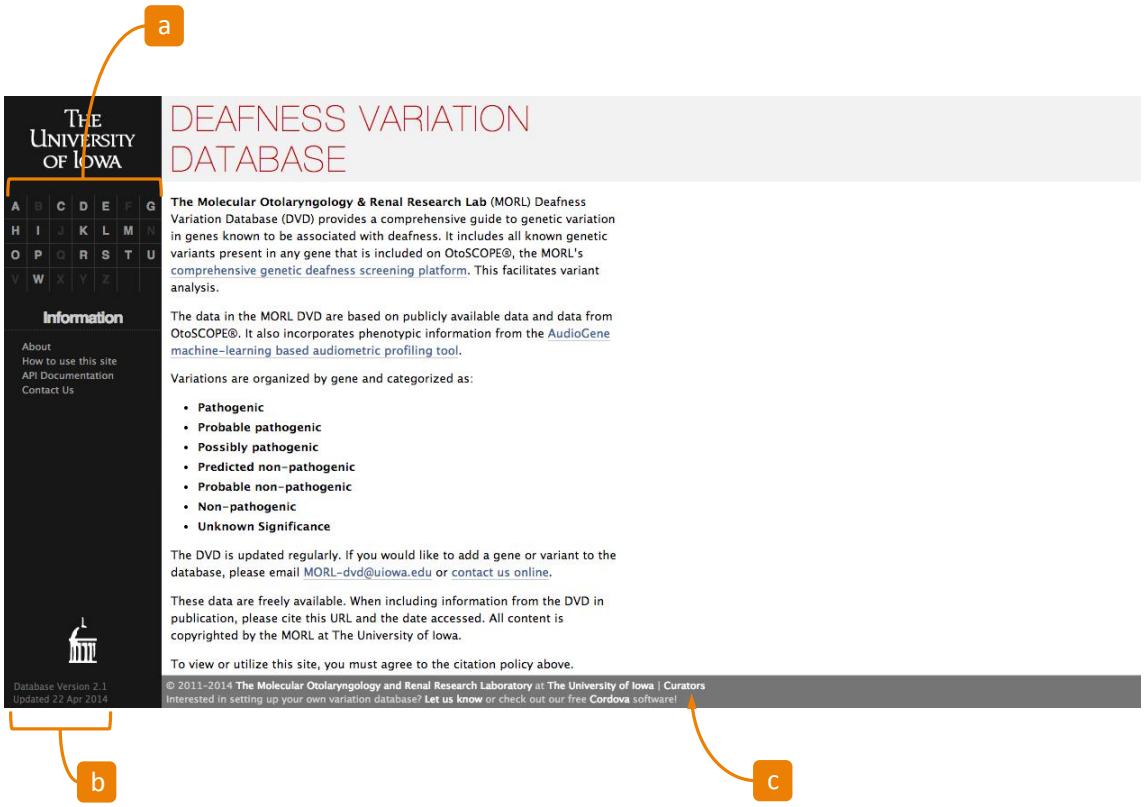

Figure 6: Homepage of the public interface. (a) The alphabetical index can be used to filter genes/loci by first letter. (b) Information is shown to indicate the current database version number and when the database was last updated. (c) A link is provided for authenticated users to log into the database for editing privileges.

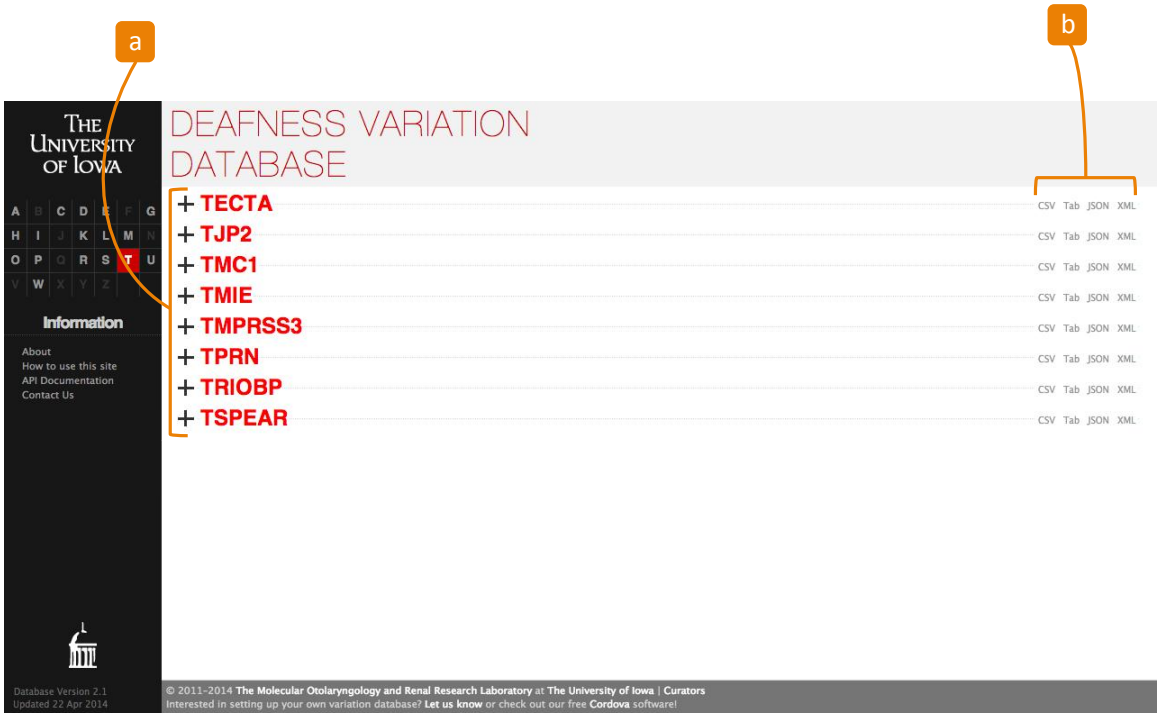

Figure 7: (a) Public interface for gene/locus results filtered by first letter and (b) a list of formats available for download.

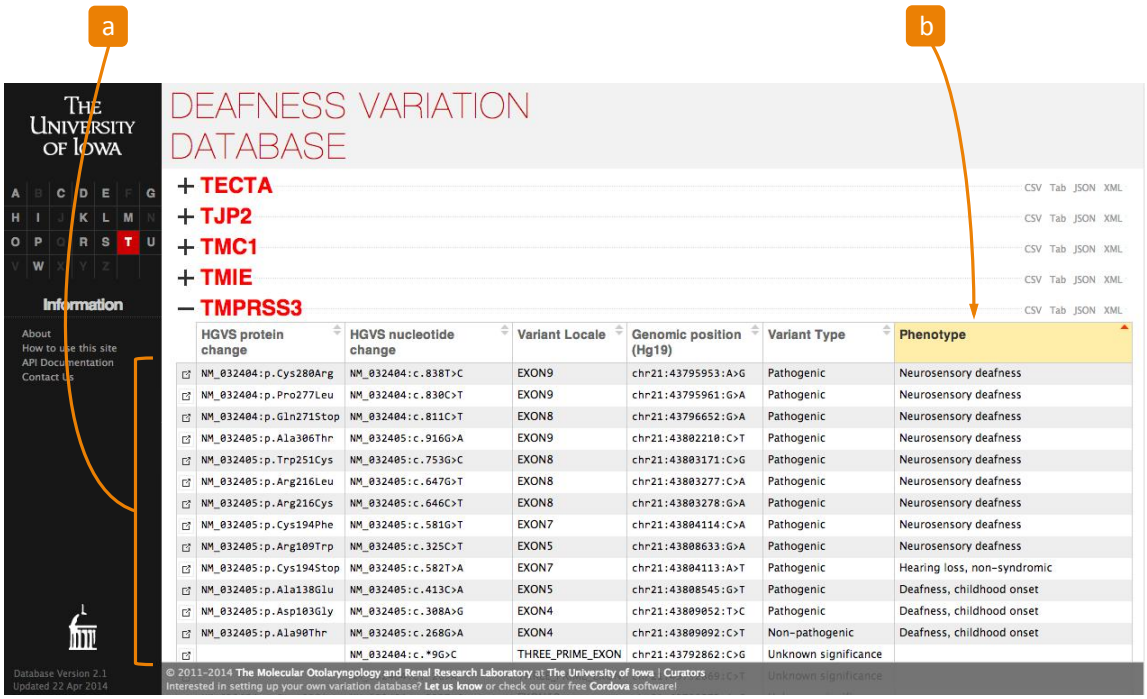

Figure 8: (a) Public interface for gene/locus results filtered by first letter and expanded to show respective variations. (b) Variations can be sorted by HGVS protein change, HGVS nucleotide change, variant locale, genomic position, variant type, or phenotype.

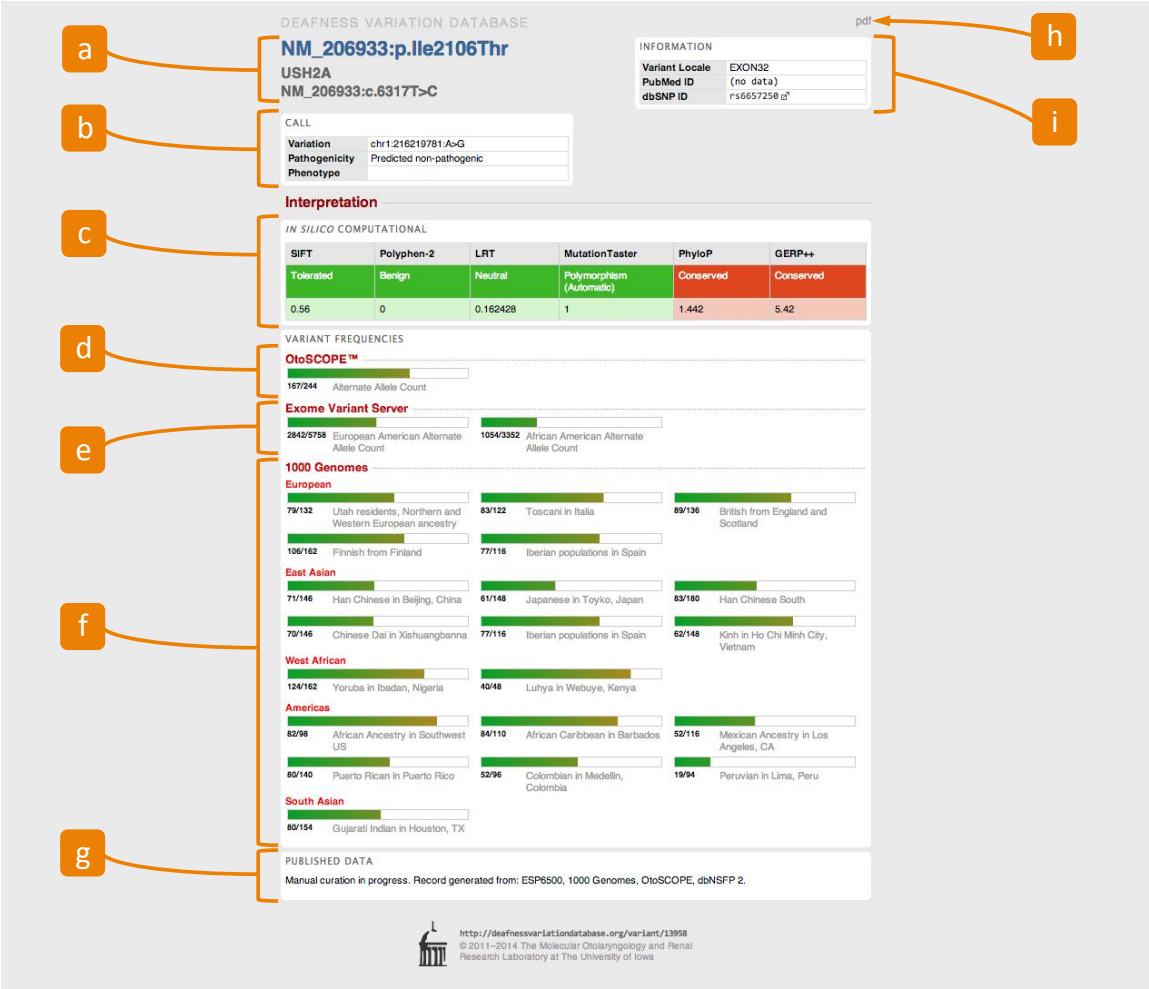

Figure 9: Public interface for a variation profile indicating (a) HGVS protein change, gene/locus, HGVS nucleotide change, (b) genomic coordinates, pathogenicity, phenotype, (c) algorithmic prediction scores, (d) OtoSCOPE® minor allele frequencies, (e) EVS minor allele frequencies, (f) 1000 Genomes minor allele frequencies, (g) curator comments, (h) link to PDF download, and (i) variant locale, PubMed ID(s), and dbSNP ID.

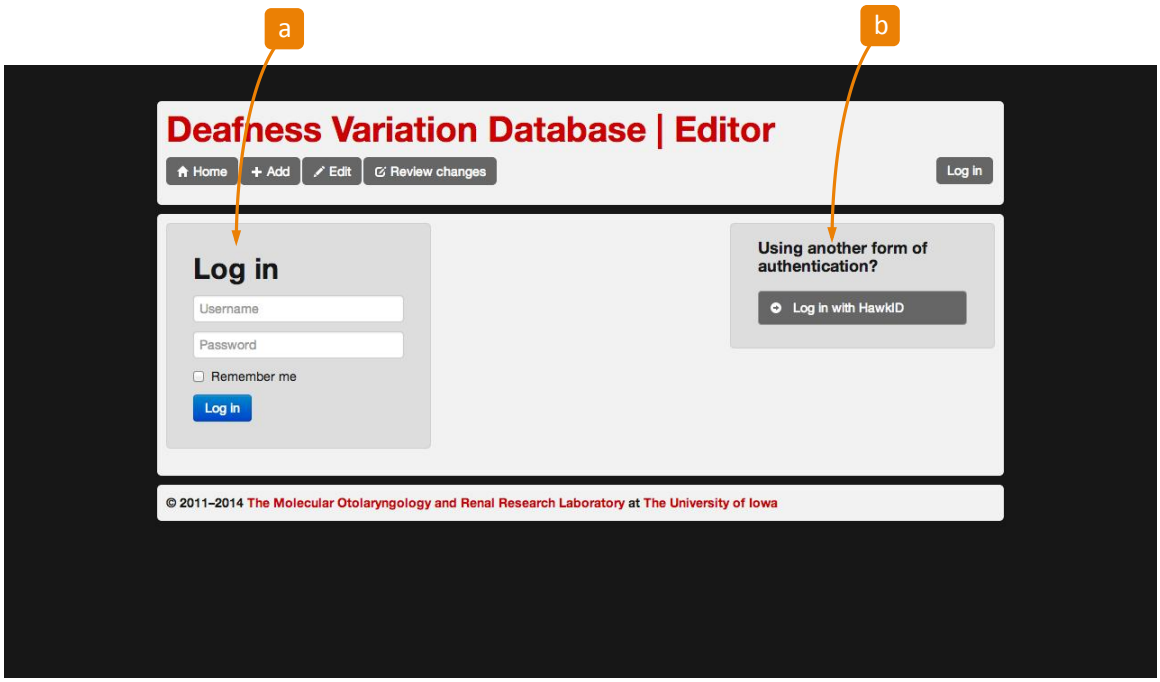

Figure 10: Public interface for logging into the private interface used for database curators and administrators. (a) A form is provided for local user authentication, and (b) a link is provided for remote user authentication.

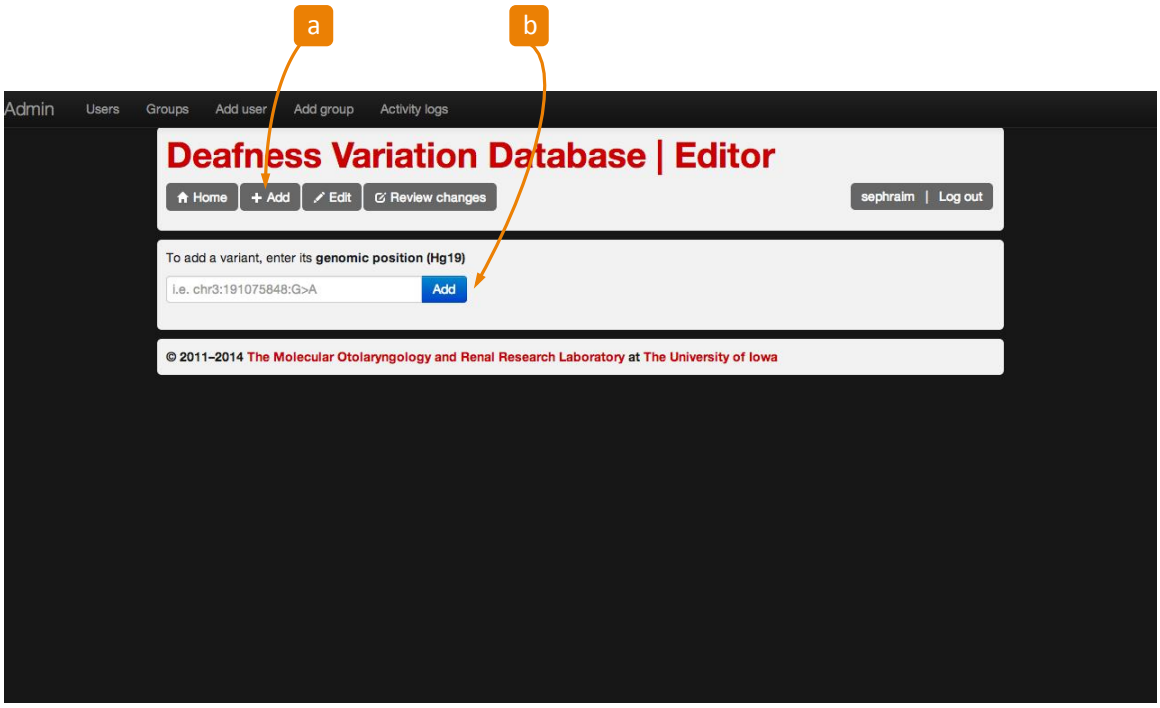

Figure 11: Private interface for database curators to add new variants. (a) A link is provided for users to get to this page. (b) A form field is provided for users to submit a new variation to the database by supplying its genomic position.

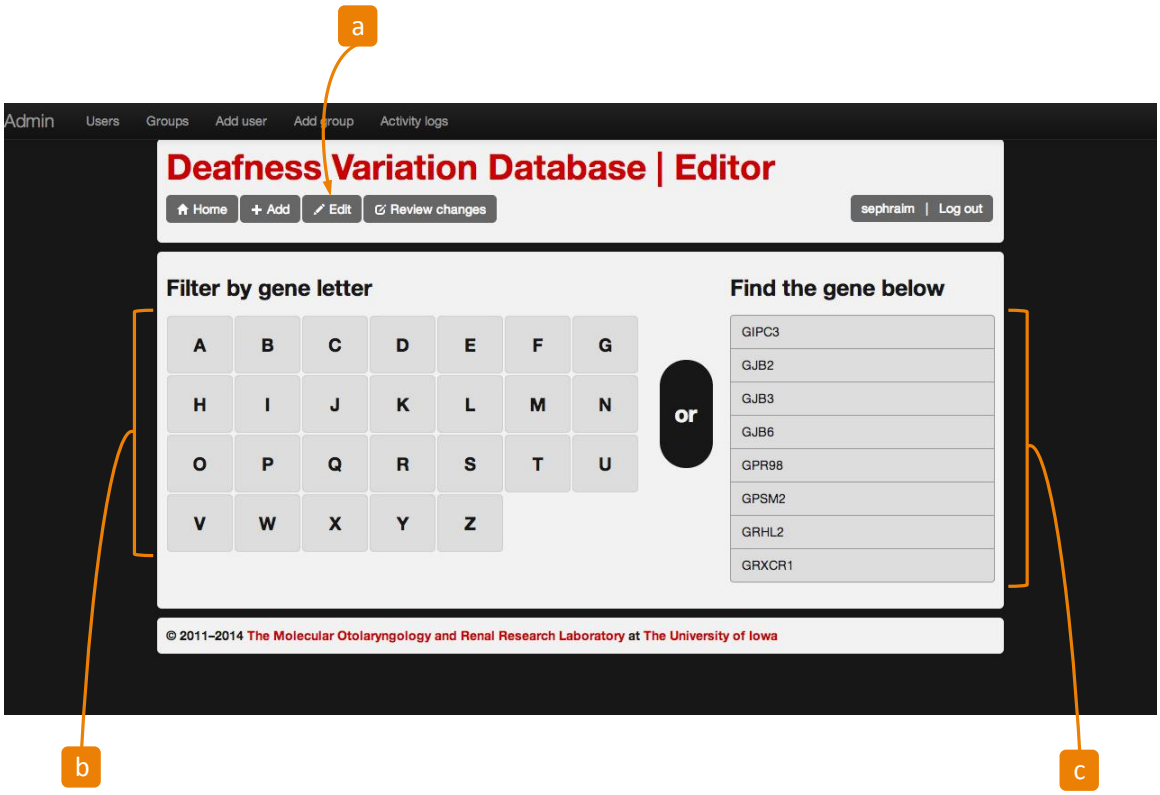

Figure 12: Private interface for filtering genes/loci by first letter to select for editing. (a) A link is provided for users to get to this page. (b) The alphabetical index can be used to bring up (c) a list of genes/loci filtered by first letter.

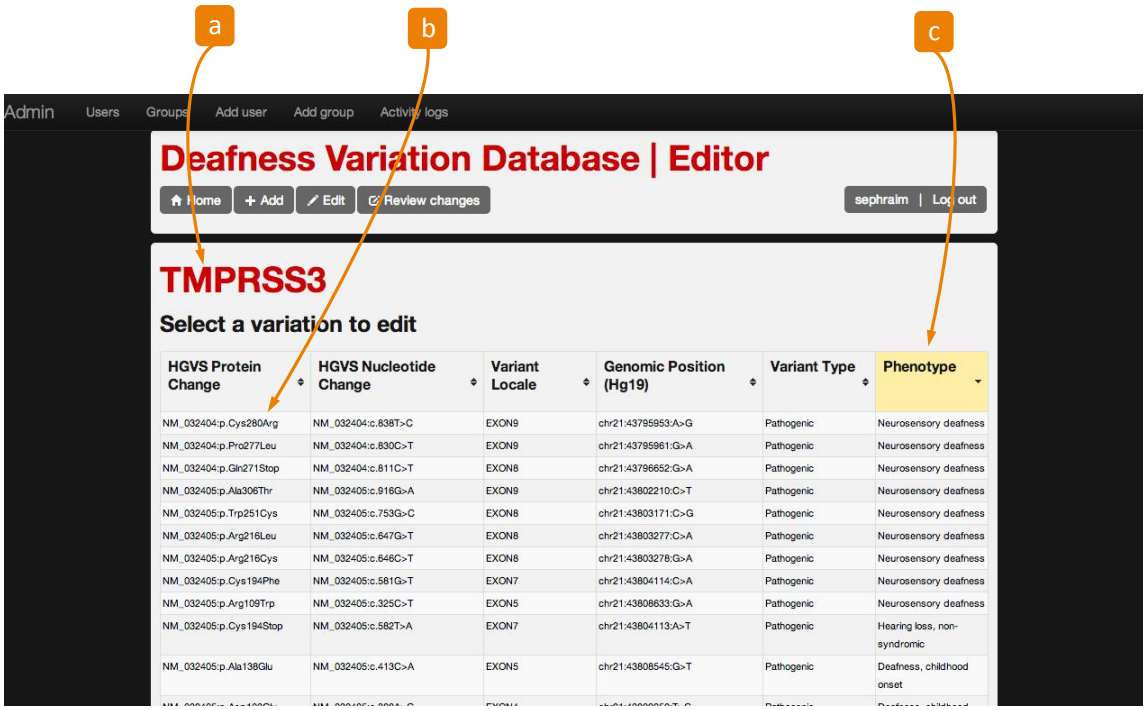

Figure 13: Private interface for filtering variations by gene/locus to select for editing. (a) The gene/loci selected from the previous page is displayed along with (b) a list of respective variations. (c) Variations can be sorted by HGVS protein change, HGVS nucleotide change, variant locale, genomic position, variant type, or phenotype.

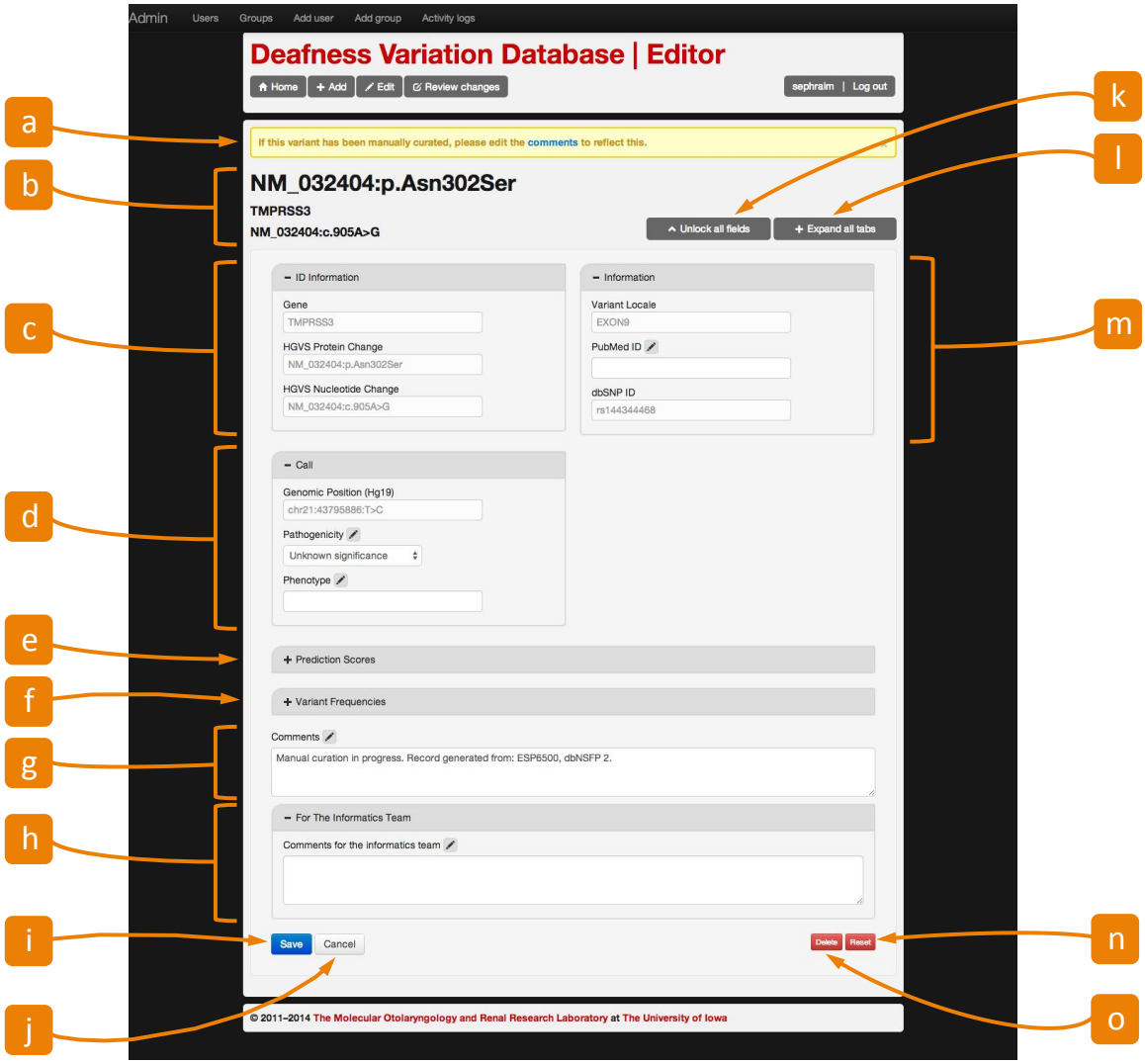

Figure 14: Private interface for editing a variation profile. (a) A message is displayed to encourage curators to edit the comments about manual curation. (b) Variation identification is displayed to clearly indicate the variation profile currently being edited. Fields are provided to edit a variation's (c) gene/locus, HGVS protein change, HGVS nucleotide change, (d) genomic position, pathogenicity, phenotype, (e) algorithmic prediction scores, (f) minor allele frequencies, (g) curator comments, (m) variant locale, PubMed ID(s), and dbSNP ID. (h) A field is provided to make private comments to other curators. Edits currently on the page can either be (i) saved to the queue or (j) canceled. (k) A button is provided to unlock all fields for editing, and another button is provided to expand all the collapsible tabs. (n) All edits currently saved to the queue can be discarded, or (o) a variation can be scheduled for deletion from the next release of the database.

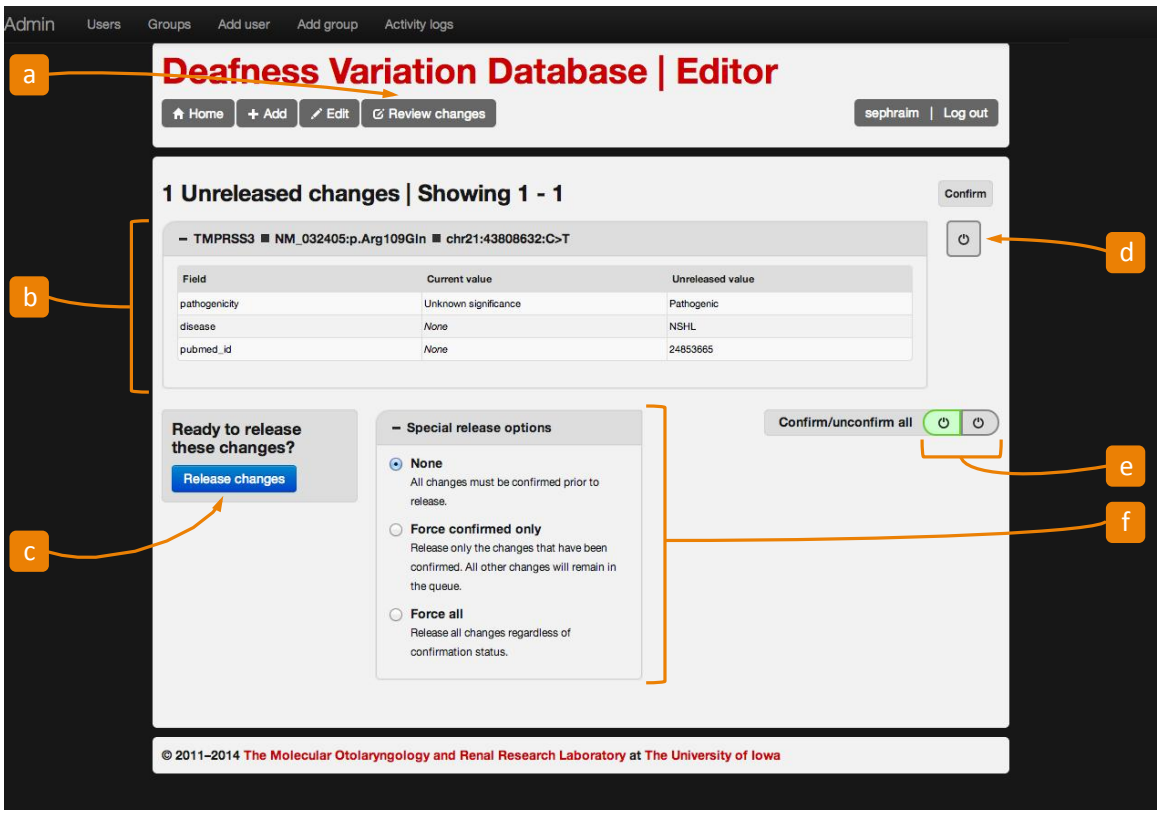

Figure 15: Private interface for reviewing unreleased changes to variation profiles. (a) A link is provided for users to get to this page. (b) A list of all edits made to a variations profile is displayed to allow the curator to see each edited field, the current (public) value of each field, and the unreleased value of each field. (c) A button is provided to allow curators to release the changes after (d) each change has been confirmed. Curators also have (e) the option to confirm or unconfirm all changes at once or select from (f) several special release options.

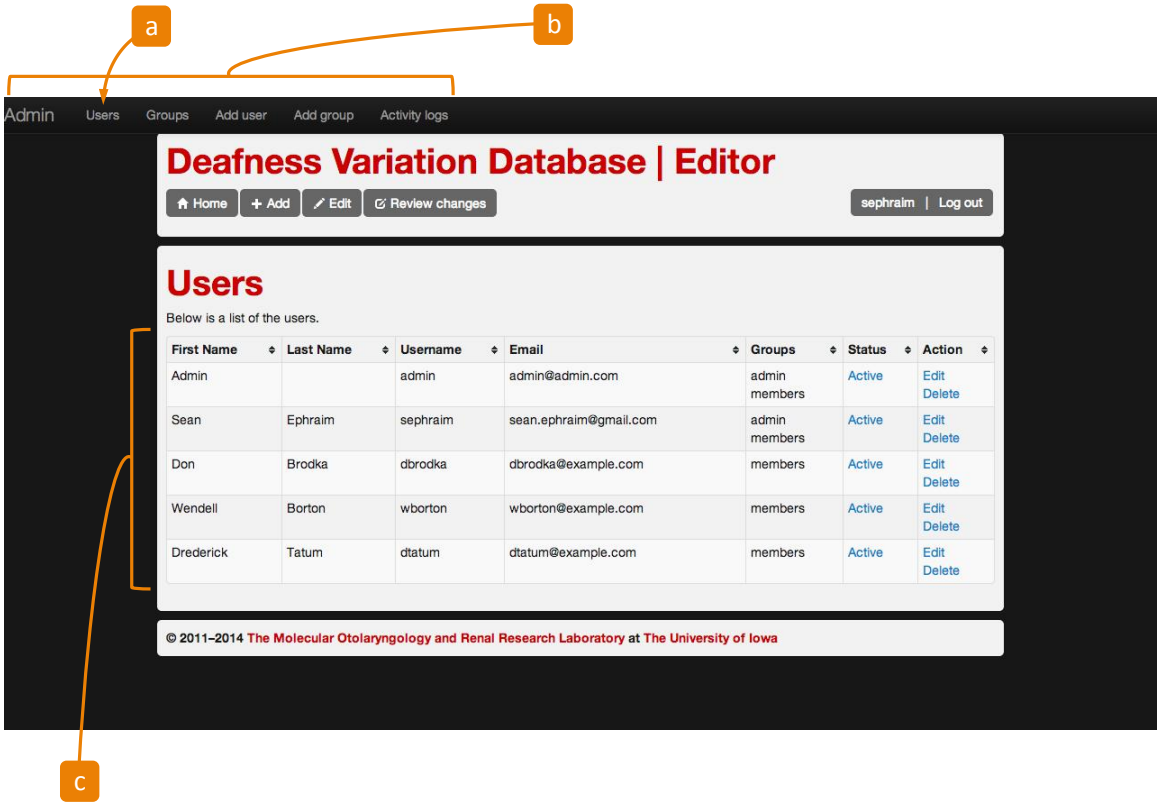

Figure 16: Private interface for administrators to review authenticated users of the variation database. (a) A link is provided for administrators to get to this page from within (b) the administrative toolbar. (c) A list of all authenticated users is displayed with options to edit user credentials, activate or deactivate user access, or delete a user altogether.

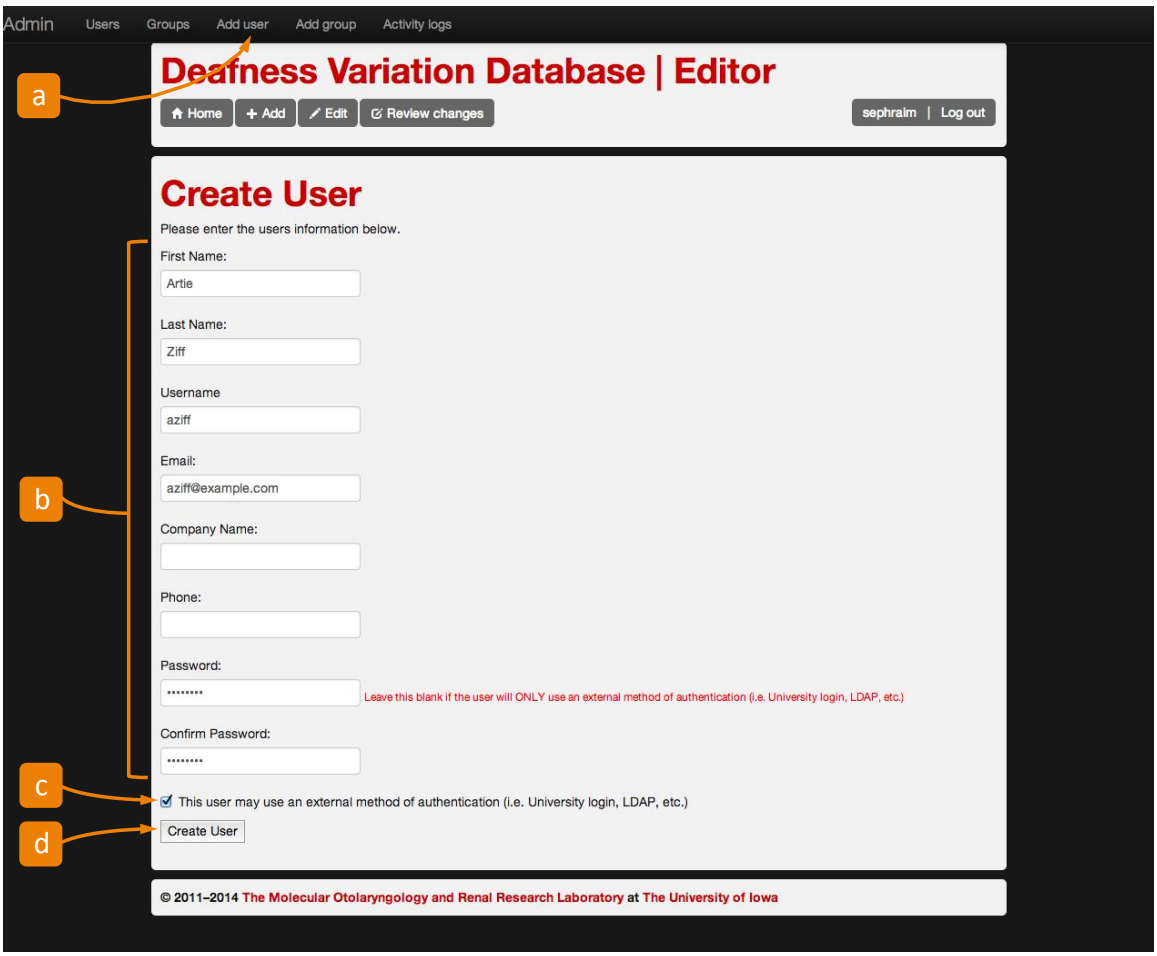

Figure 17: Private interface for administrators to create a new authorized user of the database. (a) A link is provided for administrators to get to this page from within the administrative toolbar. (b) Form fields are provided to allow administrators to edit user information, and (c) a checkbox is provided to allow administrators to indicate whether or not users will be using remote authentication. (d) A button is provided to allow administrators to create a new user after completing the new user profile.

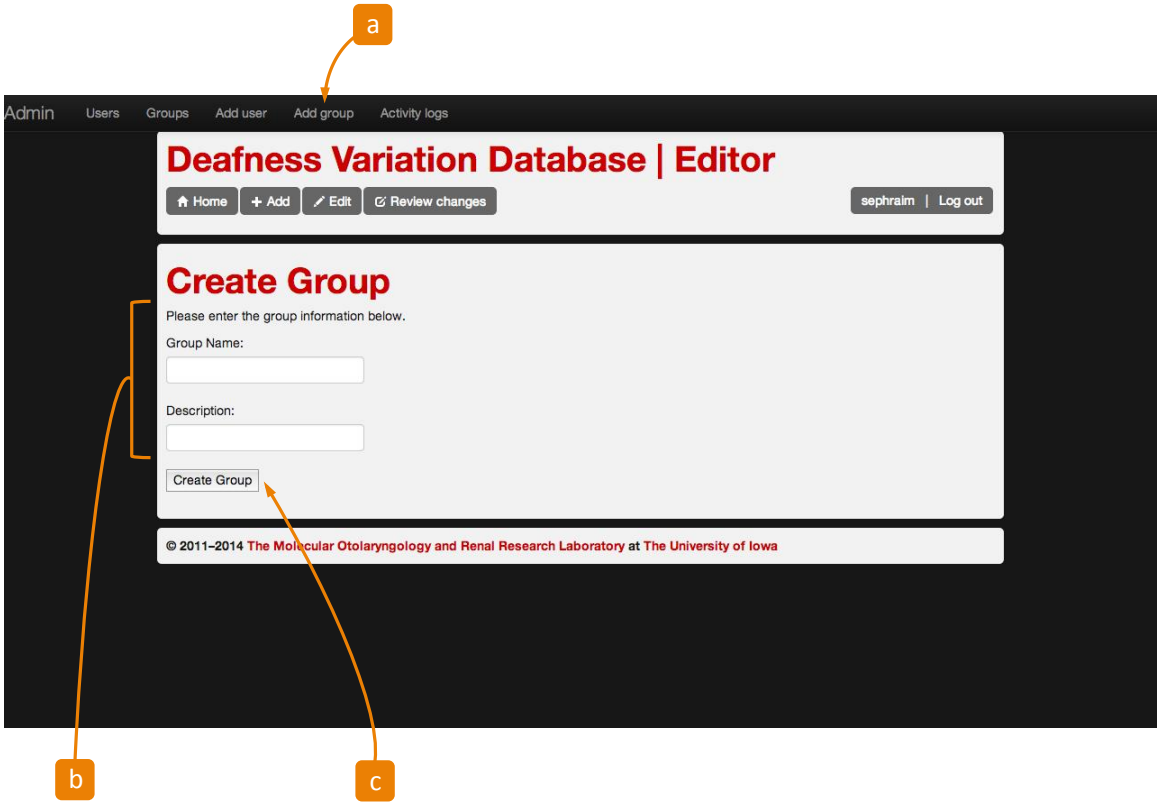

Figure 18: Private interface for administrators to create a new authorized user group of the database. (a) A link is provided for administrators to get to this page from within the administrative toolbar. (b) Form fields are provided for administrators to create a new user group with an accompanying description, and (c) button is provided for administrators to create the group.

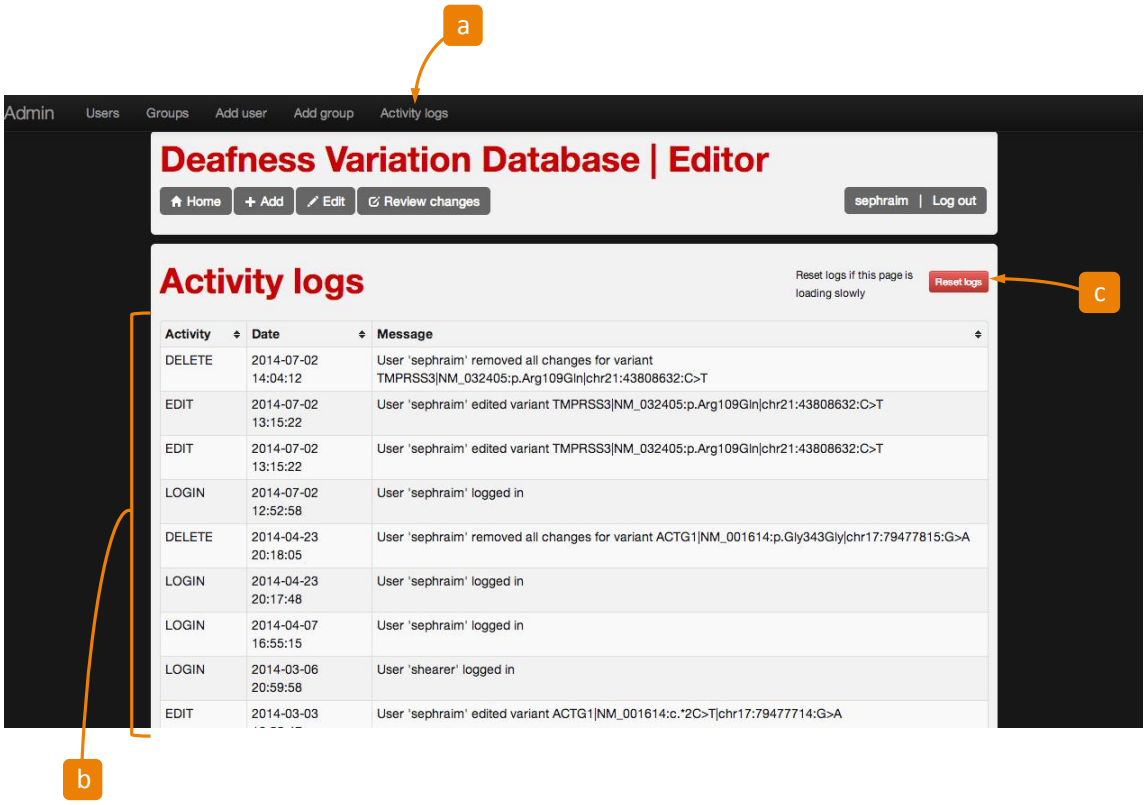

Figure 19: Private interface for administrators to view activity logs. (a) A link is provided for administrators to get to this page from within the administrative toolbar, and (b) a list of all user activity is displayed. There is also (c) an option for administrators to reset the activity logs.

4.3 Discussion

4.3.1 Interface

Cordova was designed to have two separate interfaces. The first interface is publicly available to any user. Screenshots for the public interface can be seen in Figure 6 - Figure 9 The second interface is private and only available to authenticated users. Screenshots for the private interface can be seen in Figure 10 - Figure 19.

On the public interface, information is shown at all times to indicate the current version of the database and when it was last updated. To begin browsing the database, users can utilize the alphabetical interface to filter disease-associated genes and loci by

their first letter. Clicking on a letter will bring up a list of all genes and loci currently available within the database. A user can then click on any of the listed genes or loci to reveal a list of all associated variants and sort them by HGVS protein change, HGVS nucleotide change, variant locale, genomic position, variant type, or phenotype. Alternatively, the user can click on the CSV, tab, JSON, XML links to download the variant information in any of these formats. Clicking a variant will reveal its annotation profile as seen in Figure 9, which includes data for the gene or locus, genomic coordinates, HGVS nucleotide change, HGVS protein change, variant locale, pathogenicity, phenotype, algorithmic prediction scores, OtoSCOPE® MAFs, EVS MAFs, 1000 genomes MAFs, PubMed ID, dbSNP ID, and curator comments. Each profile also includes a link to download the profile in PDF format.

Users can click the "Curators" link at the bottom of the public interface to log into the private interface as shown in Figure 10. Two options are available for users to securely log in: an option to provide authentication credentials stored in the local database and an option to authenticate remotely through the University of Iowa servers. Once logged in, users can create new variant profiles by simply supplying genomic coordinates, reference allele(s), and alternate alleles(s) in the form shown in Figure 11. Users can also edit existing variants in the database by utilizing the alphabetical index shown in Figure 12 to select a gene or locus and then selecting a respective variant as shown in Figure 13. Users will then be taken to a variant profile page as shown in Figure 14 in which they can edit associated fields. For quality control, users can review and confirm any change that has been made using the interface shown in Figure 15 before choosing to release them to the public. Special options are available to allow curators to release only changes that have been confirmed (while withholding all others) or forcing the release of all changes regardless of their confirmation status.

Administrators are able to view a list of all authenticated users of the database using the interface shown in Figure 16. This list provides options to edit user credentials,

activate or deactivate user access, or delete a user altogether. Additionally, administrators can create a new authenticated user or user group using the interfaces shown in Figure 17 and Figure 18, respectively. Lastly, administrators can view all important activity logs for authenticated users as shown in Figure 19. This includes logs for when a user logs in, logs out, creates a new variant, edits an existing variant, resets variant edits, or releases a new version of the database.

4.3.2 Features and data management

Cordova is written in PHP, built on the popular CodeIgniter 2 web application framework, and utilizes 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.

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 dbNSFP 2 (Liu *et al.*, 2013), EVS (2013), 1000 Genomes (2012) and OtoSCOPE® (Shearer *et al.*, 2013). Each MAF source can be configured for

inclusion or exclusion. For example, the OtoSCOPE® source was excluded from the VVD. Auto-populated fields 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 in order 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 utilizes 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 upon review. Options are available either to release 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), LRT (Chun and Fay, 2009), MutationTaster (Schwarz *et al.*, 2010), phyloP (Siepel *et al.*, 2006), GERP++ (Davydov *et al.*, 2010). Prior to 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.

Table 5: Comparison of notable features and specifications for Cordova and LOVD.

	Cordova	**LOVD**
Computational 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

4.3.3 Architectural pattern

As mentioned before, the CodeIgniter 2 web application framework was selected to implement a genetic variation database manager. This framework uses the MVC architectural pattern, which is a popular choice for a software frameworks The purpose of this architecture is to separate the infrastructure into three separate layers in order to organize the code in a well-understood manner. The developers of LOVD did not use a

web application framework, and the LOVD documentation does not state the nature of its architectural pattern. After examining the source code, it is not clear as to whether LOVD was built using a known architectural pattern or if the developers chose to implement their own. It is, however, clear that LOVD does not use the MVC architecture.

4.3.4 Database abstraction

CodeIgniter features a database abstraction layer for increased flexibility of integration with various database languages including MySQL, MySQLi, MS SQL, Postgres, Oracle, SQLite, and ODBC. This essentially allows a programmer to write database queries using PHP code instead of a vendor's specific database language. In turn, this also allows the queries to become database-agnostic, meaning they're compatible with several database management systems instead of just one. This will often reduce the amount of code needed for maximum database compatibility. Cordova utilizes CodeIgniter's database abstraction layer for most of its database queries, meaning that Cordova is mostly compatible with all of CodeIgniter's supported database languages. However, some queries are quite complex and cannot be handled by the database abstraction layer alone. These queries were therefore written in MySQL. This means Cordova is known only to be fully compatible with MySQL, but in some cases a MySQL query can be identical to that of another database language. Further testing would need to be conducted to determine the compatibility with CodeIgniter's other supported database languages.

4.3.5 Annotation pipeline

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 dbNSFP 2. It uses compressed versions of annotation files to reduce their

total footprint by one-third and runs approximately eight times faster than dbNSFP 2's native querying application for large sets of annotations.

tabix is an open source command line tool written in the C programming language. It is distributed with both the source code and an executable binary. In order to be able to use tabix with Kafeen, Ruby "binding" was written to effectively execute the tabix binary using only Ruby.

In addition, configuration file was created to allow users to set some of the default behavior of Kafeen. Such configurations include allowing the user to alter the labels assigned for clinical significance predictions or changing the minimum number of prediction scores needed to assign an overall prediction. To use Kafeen as a standalone command line tool, a user can put the name of the Hg19 genomic position (e.g. chr1:216251520:T>G) into a file, run the annotation command, and view the annotations in the output file and any errors in the log file. Full usage documentation can be found at https://github.com/clcg/kafeen.

4.3.6 dbNSFP versioning

Cordova was originally designed to use prediction scores from dbNSFP 1. Later on, it was redesigned to implement prediction scores from dbNSFP 2. However, there are a few important differences between dbNSFP 1 and dbNSFP 2 that are worth noting.

dbNSFP 1 contains SIFT and phyloP scores that have been converted to a scale that is specific to dbNSFP 1. These scores are not the original scores that SIFT and phyloP actually produce, but they are used to come up with a clean prediction (i.e. "Damaging", "Benign", etc.) The columns in dbNSFP 1 for these converted SIFT and phyloP scores are called "SIFT_score" and "PhyloP_score", respectively. This is rather misleading, and realistically these columns should be called "SIFT_score_converted", and "PhyloP_score_converted". With these scores, variants are considered damaging if they have a score greater than 0.95.

dbNSFP 2.0 (released February 2013) contains only the raw scores from SIFT and phyloP instead of the aforementioned converted scores. The columns have the same names as dbNSFP 1 ("SIFT_score" and "PhyloP_score") even though they contain different data compared to dbNSFP 1. The converted scores and the predictions for SIFT and phyloP were completely removed from this version. With these scores, variants are considered damaging if they have a SIFT score less than 0.05 or a phyloP score greater than 1.

In order to cope with the missing predictions from dbNSFP 2.0, Cordova computes predictions on-the-fly. However, this posed a problem when data from dbNSFP 1 was being used for a computation that required data from dbNSFP 2. For example, a SIFT score for the GJB2 variant p.M195V has a value of 0.96 in dbNSFP 1 and an equivalent value of 0.02 in dbNSFP 2 as shown in Table 6:. Likewise, the same variant has a phyloP score of 0.998 in dbNSFP 1 and a value of 2.279 in dbNSFP 2 as shown in Table 6:.

Table 6: SIFT and phyloP prediction scores from dbNSFP 1 and equivalent scores from dbNSFP 2 for the GJB2 variant p.M195V.

	dbNSFP 1	**dbNSFP 2**
SIFT	0.960	0.020
phyloP	0.998	2.279

It wasn't until dbNSFP 2.1 (released October 2013) that the converted scores and prediction for SIFT were included. However, there are still no converted scores or predictions for phyloP.

4.3.7 Auto-installation of third-party dependencies

The auto-installation feature for fetching and installing the dompdf 0.5 and pChart 2 libraries were a design choice chosen for licensing issues. The decision was made to distribute the source code for Cordova under the MIT license. An unrestrictive license would allow users to freely utilize Cordova and maximize the opportunity for developers to expand it. The dompdf 0.5 and pChart 2 libraries are distributed under the LGPL and GPL licenses, respectively, both of which have more restrictive limitations than the MIT license. Distributing these libraries directly with the Cordova software would inherently make Cordova restricted to the GPL license agreement as well.

4.3.8 Technology decisions

With the decision to generalize the Cordova software to allow curation and dissemination for any set of variants, the software was implemented using a web application framework. Ideally, an appropriate web application framework would be: 1) stable for long-term use; 2) extensible for the incorporation of new features; and 3) well-documented for developers to easily become acquainted with the capabilities. The first prototype of Cordova was implemented using the Drupal content management system.

Content management systems generally offer a large number of out-of-the-box features to reduce the amount of code a developer must write. However, these solutions are usually quite rigid and should typically be used for developing web applications that can benefit from utilizing many of these available features. While Drupal remains a popular choice for web development, it was not flexible enough to accommodate the features we ultimately wanted to achieve. The second version was from scratch in PHP without the use of a content management system or web application framework. Although this application served the specific needs of the DVD, it could not be easily extended to serve as a general variation database. In addition, the documentation was quite minimal, making it difficult for new developers to quickly become familiar the infrastructure. Moreover, the application did not contain a graphical user interface for interacting with the database in order to edit records.

For the third and current phase, Cordova was implemented using the CodeIgniter 2 web application framework. In addition to PHP, other popular web programming languages such as Ruby and Java have their own respective frameworks as well. However, because the second version of Cordova was written in PHP, we decided to choose a web application framework for PHP. The second version contained many important stable and well-tested features such as a robust API, which were all written in PHP, it was desirable to choose a framework that could allow the import many of these existing features without whole-scale reimplementation. PHP frameworks other than CodeIgniter 2 were also considered such as CakePHP, Yii, and Laravel. Each of these frameworks has comparable features, but CodeIgniter 2 was selected for its flexibility, ease of use, built-in features and thorough documentation.

A design goal for Kafeen was the desire for an annotation pipeline that could easily be integrated into a web infrastructure as well as be used as a standalone command line tool. The original prototype of Kafeen was written in PHP. While PHP is a well-accepted web programming language, it is typically not used for developing command

line tools. Additionally, the annotation involved heavy string processing tasks that proved to be quite cumbersome using PHP. Therefore the production release of Kafeen was developed using Ruby, a mature object-oriented programming language. In addition to developing command line tools, Ruby is also a standard web programming language. Ruby provides a large standard library of methods including string processing. The BioRuby library for bioinformatics-based modules is also available to Ruby developers, although the current version of Kafeen does not take advantage of it.

Cordova and Kafeen are written in two separate programming languages, PHP and Ruby, respectively. Though not necessary, it would be possible to port Cordova from PHP to Ruby for easier maintainability of both components. Python and Java are also mature object-oriented programming languages used for the development of web applications and command line tools. However, Ruby was ultimately chosen for its robust web application framework, Ruby on Rails. Python and Java were also considered as an alternative to Ruby. While each of these languages provides several web application framework choices, Ruby on Rails powers many high-traffic websites such as GitHub (https://github.com), Bloomberg (http://www.bloomberg.com), Hulu (http://www.hulu.com), and Twitter (https://twitter.com). Should Cordova ever be rewritten in Ruby, support of Kafeen and Cordova would be reduced to one common framework (Ruby on Rails).

The tabix indexing tool was implemented into the Kafeen annotation pipeline because of its ability to quickly retrieve annotations from large data files. The original prototype of Kafeen used the native software packaged with dbNSFP 2 to retrieve annotations from its data files. While this was a working solution, it also posed three problems: 1) retrieving annotations was quite slow; 2) the software could only query the uncompressed dbNSFP data files, which were quite large; 3) the software was specific to dbNSFP, and could not be used for other data files. By introducing tabix into the pipeline, each of these problems was solved. Firstly, tabix proved to query the data files

approximately eight times faster than the native querying software for dbNSFP 2. Second, tabix was able to query the compressed versions of data files, reducing the overall footprint of Kafeen's data files from approximately 30GB to 10GB. All data files were compressed using the bgzip compression tool packaged with tabix. Lastly, tabix was a uniform solution for querying all of Kafeen's data files including dbNSFP 2, 1000 Genomes, EVS, and OtoSCOPE®.

4.3.9 Cordova installation requirements

Cordova requires a Linux/Unix-based operating system (e.g. Ubuntu, CentOS, Mac, etc.), an Apache web server (enabled for PHP), PHP 5.3.0 or greater (php-xml extension must be enabled for PDF generation), MySQL 5.0.95 or greater, and Sendmail (needed for email service). Cordova was developed using only these specifications, but it possible that Cordova is compatible with other technologies as well.

4.3.10 Kafeen installation requirements

Kafeen can be installed on Windows or a Linux/Unix-based operating system (e.g. Ubuntu, CentOS, Mac, etc.). It requires Ruby 1.9 or greater and the Java SE Runtime Environment 1.7 or greater. In addition, approximately 10GB of disk space is required for the annotation data files.

4.3.11 Availability

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

CHAPTER 5: CONCLUSION

HGMD contains genetic variation data from over 6000 genes with variation records totaling over 148,000. HGMD has clearly indicated that their policy is to enter any variation into the database that has been associated with disease, even if the functional significance is unclear (Stenson *et al.*, 2014). While it is important to avoid the exclusion of variations that could possibly be linked to disease, this introduces a high false-positive rate for disease-causing mutations. HGMD denotes all such questionable mutations with a question mark. However, our findings indicate inconsistencies, as seen in Table 3.

Containing data from only 83 genes, the DVD serves as a much smaller variation database than HGMD. However, this allows curators to perform focused analyses to enhance the quality of the data. While HGMD serves as a very valuable resource for collating variations that have been associated with disease in some way, it is difficult for such a large and ever-growing number of variant records to stay up-to-date. This creates an increasing need for smaller LSDBs.

Cordova was created to provide a set of tools for curators and clinicians to easily create their own LSDB. Such databases have a narrower, refined focus, making them easier to manage in order to increase the quality of the data. LOVD is currently the only other published software available for setting up a LSDB. Prior to creating Cordova, we looked into expanding LOVD to add missing features such as data versioning and computationally predicting clinical significance. However, LOVD provides no developer documentation, making it difficult for new developers to learn the system and add new features. For this reason, we opted to create an entirely new, fully documented system with unrestrictive licensing to allow any developers to create and update features for long-term sustainability.

APPENDIX

Table A-1: The phenotypic labels for deafness-associated variants from HGMD and/or ClinVar and their respective standardized names for DVD.

Phenotypic label from HGMD and/or ClinVar	**Phenotypic label for DVD**
AllHighlyPenetrant; Deafness, autosomal recessive 9	NSHL
AllHighlyPenetrant; Retinitis pigmentosa-deafness syndrome	Usher Syndrome
Auditory neuropathy	Auditory neuropathy
Auditory neuropathy spectrum disorder	Auditory neuropathy
Auditory neuropathy, autosomal recessive, 1	Auditory neuropathy
Auditory neuropathy, autosomal recessive, 1; Deafness, autosomal recessive 9	Auditory neuropathy
Auditory neuropathy, temperature-sensitive	Auditory neuropathy
Cardiomyopathy, hypertrophic with deafness	HCM with deafness
Deafness	NSHL
Deafness and brachycardia	Long QT syndrome
Deafness and palmoplantar hyperkeratosis	Deafness and palmoplantar hyperkeratosis
Deafness and palmoplantar keratoderma	Deafness and palmoplantar keratoderma
Deafness and vestibular areflexia	Deafness and vestibular areflexia
Deafness, autosomal dominant	NSHL
Deafness, autosomal dominant 11	NSHL
Deafness, autosomal dominant 12	NSHL
Deafness, autosomal dominant 13	NSHL
Deafness, autosomal dominant 15	NSHL
Deafness, autosomal dominant 17	NSHL
Deafness, autosomal dominant 2	NSHL
Deafness, autosomal dominant 20	NSHL
Deafness, autosomal dominant 22	NSHL
Deafness, autosomal dominant 23; Branchiootic syndrome 3	NSHL/BOR
Deafness, autosomal dominant 25	NSHL
Deafness, autosomal dominant 28	NSHL
Deafness, autosomal dominant 2b	NSHL
Deafness, autosomal dominant 3	NSHL

Table A-1: Continued.

Deafness, autosomal dominant 36	NSHL
Deafness, autosomal dominant 3a	NSHL
Deafness, autosomal dominant 3a; Deafness, autosomal recessive 1A	NSHL
Deafness, autosomal dominant 3a; Hereditary hearing loss and deafness	NSHL
Deafness, autosomal dominant 4	NSHL
Deafness, autosomal dominant 4; AllHighlyPenetrant	NSHL
Deafness, autosomal dominant 48	NSHL
Deafness, autosomal dominant 48; AllHighlyPenetrant	NSHL
Deafness, autosomal dominant 4b	NSHL
Deafness, autosomal dominant 64	NSHL
Deafness, autosomal dominant 8	NSHL
Deafness, autosomal dominant nonsyndromic	NSHL
Deafness, autosomal dominant nonsyndromic sensorineural 17; MYH9 related disorders	NSHL/MYH9 related diseases
Deafness, autosomal dominant nonsyndromic sensorineural 39, with dentinogenesis imperfecta 1	NSHL/dentinogenesis imperfecta
Deafness, autosomal dominant nonsyndromic sensorineural 39, with dentinogenesis imperfecta 1; Dentinogenesis imperfecta - Shield's type II; Dentinogenesis imperfecta - Shield's type III	NSHL/dentinogenesis imperfecta
Deafness, autosomal recessive	NSHL
Deafness, autosomal recessive 1	NSHL
Deafness, autosomal recessive 12	NSHL
Deafness, autosomal recessive 12, modifier of	NSHL -- modifier
Deafness, autosomal recessive 12; Hereditary hearing loss and deafness; Retinitis pigmentosa-deafness syndrome	NSHL/Usher syndrome
Deafness, autosomal recessive 15	NSHL
Deafness, autosomal recessive 18; AllHighlyPenetrant	NSHL
Deafness, autosomal recessive 18b	NSHL
Deafness, autosomal recessive 1A	NSHL
Deafness, autosomal recessive 1A; AllHighlyPenetrant	NSHL
Deafness, autosomal recessive 1A; Deafness, digenic, GJB2/GJB3; Hereditary hearing loss and deafness	NSHL
Deafness, autosomal recessive 1A; Deafness, digenic, GJB2/GJB6; Hereditary hearing loss and deafness	NSHL
Deafness, autosomal recessive 1A; Hereditary hearing loss and deafness	NSHL

Table A-1: Continued.

Deafness, autosomal recessive 1A; Hereditary hearing loss and deafness; Deafness, autosomal dominant 3a	NSHL
Deafness, autosomal recessive 2	NSHL
Deafness, autosomal recessive 2; Usher syndrome, type 1B	NSHL/Usher syndrome
Deafness, autosomal recessive 21	NSHL
Deafness, autosomal recessive 23	NSHL
Deafness, autosomal recessive 24	NSHL
Deafness, autosomal recessive 25	NSHL
Deafness, autosomal recessive 28	NSHL
Deafness, autosomal recessive 29	NSHL
Deafness, autosomal recessive 3	NSHL
Deafness, autosomal recessive 30	NSHL
Deafness, autosomal recessive 31	NSHL
Deafness, autosomal recessive 35	NSHL
Deafness, autosomal recessive 37	NSHL
Deafness, autosomal recessive 39	NSHL
Deafness, autosomal recessive 42	NSHL
Deafness, autosomal recessive 48	NSHL
Deafness, autosomal recessive 53	NSHL
Deafness, autosomal recessive 59	NSHL
Deafness, autosomal recessive 6	NSHL
Deafness, autosomal recessive 63	NSHL
Deafness, autosomal recessive 67	NSHL
Deafness, autosomal recessive 7	NSHL
Deafness, autosomal recessive 7; Hereditary hearing loss and deafness	NSHL
Deafness, autosomal recessive 70	NSHL
Deafness, autosomal recessive 74	NSHL
Deafness, autosomal recessive 77	NSHL
Deafness, autosomal recessive 79	NSHL
Deafness, autosomal recessive 8	NSHL
Deafness, autosomal recessive 84b	NSHL
Deafness, autosomal recessive 9	NSHL
Deafness, autosomal recessive 9; AllHighlyPenetrant	NSHL

Table A-1: Continued.

Deafness, autosomal recessive 9; Auditory neuropathy, autosomal recessive, 1	NSHL/Auditory Neuropathy
Deafness, autosomal recessive 9; Hereditary hearing loss and deafness	NSHL
Deafness, autosomal recessive 9; Malignant melanoma	NSHL
Deafness, autosomal recessive 91	NSHL
Deafness, bilateral, with inner ear malformation	NSHL with EVA
Deafness, childhood onset	NSHL
Deafness, digenic, GJB2/GJB3	NSHL
Deafness, dominant progressive	NSHL
Deafness, neurosensory, autosomal recessive 49	NSHL
Deafness, non-syndromic	NSHL
Deafness, non-syndromic, autosomal dominant	NSHL
Deafness, non-syndromic, autosomal dominant 1	NSHL
Deafness, non-syndromic, autosomal recessive	NSHL
Deafness, nonsyndromic	NSHL
Deafness, nonsyndromic sensorineural	NSHL
Deafness, nonsyndromic sensorineural 25	NSHL
Deafness, sensorineural non-syndromic 11	NSHL
Deafness, unilateral	NSHL
Deafness, without vestibular involvement, autosomal dominant	NSHL
Deafness, X-linked 1	XLNSHL
Deafness, X-linked 2	XLNSHL
Deafness, X-linked 4	XLNSHL
DFNA 2 Nonsyndromic Hearing Loss	NSHL
DFNA36 hearing loss, association with	NSHL
DFNB7/B11 deafness	NSHL
Diabetes mellitus AND insipidus with optic atrophy AND deafness	Wolfram syndrome
Diabetes, type 1, and sensorineural hearing loss	Wolfram syndrome
Enlarged vestibular aqueduct	NSHL with EVA
Enlarged vestibular aqueduct & Mondini deformity	NSHL with EVA/Mondini
Enlarged vestibular aqueduct & Mondini dysplasia	NSHL with EVA/Mondini
Enlarged vestibular aqueduct & vestibular dilatation	NSHL with EVA/vestibular dilatation
Enlarged vestibular aqueduct syndrome	NSHL with EVA

Table A-1: Continued.

Enlarged vestibular aqueduct syndrome; Hereditary hearing loss and deafness; Pendred's syndrome	NSHL with EVA/Pendred syndrome
Enlarged vestibular aqueduct syndrome; Pendred's syndrome	EVA/Pendred syndrome
Enlarged vestibular aqueduct syndrome; Pendred's syndrome; Hereditary hearing loss and deafness	NSHL with EVA/Pendred syndrome
Epilepsy, ataxia, sensorineural deafness and tubulopathy (EAST syndrome)	Epilepsy, ataxia, sensorineural deafness and tubulopathy (EAST syndrome)
Hearing impairment	NSHL
Hearing impairment, autosomal recessive	NSHL
Hearing impairment, bilateral sensorineural	NSHL
Hearing impairment, nonsyndromic	NSHL
Hearing impairment, nonsyndromic, autosomal recessive	NSHL
Hearing loss	NSHL
Hearing loss with dilation of vestibular aqueduct	NSHL with EVA
Hearing loss, autosomal dominant	NSHL
Hearing loss, digenic non-syndromic	NSHL
Hearing loss, non-syndromic	NSHL
Hearing loss, non-syndromic sensorineural	NSHL
Hearing loss, non-syndromic, autosomal dominant	NSHL
Hearing loss, non-syndromic, autosomal recessive	NSHL
Hearing loss, non-syndromic, sensorineural	NSHL
Hearing loss, progressive	NSHL
Hearing loss, unilateral	NSHL
Hearing loss, X-linked nonsyndromic	XLNSHL
Hereditary hearing loss and deafness	NSHL
Hereditary hearing loss and deafness; Deafness, autosomal recessive 1A	NSHL
Hereditary hearing loss and deafness; Deafness, autosomal recessive 9	NSHL
Hereditary hearing loss and deafness; Keratitis-ichthyosis-deafness syndrome, autosomal dominant	Keratitis-ichthyosis-deafness syndrome
Hereditary hearing loss and deafness; Pendred's syndrome	NSHL/Pendred syndrome
Hereditary hearing loss and deafness; Retinitis pigmentosa-deafness syndrome	NSHL/Usher syndrome
Jervell and Lange-Nielsen syndrome	JLNS
Jervell and Lange-Nielsen syndrome; Long QT syndrome, LQT1 subtype	JLNS/Long QT syndrome

Table A-1: Continued.

KCNQ1-related Jervell and Lange-Nielsen syndrome; Long QT syndrome 1/2, digenic; KCNQ1-related acquired long QT syndrome; Long QT syndrome, LQT1 subtype	JLNS/Long QT syndrome
Keratitis-ichthyosis-deafness syndrome	Keratitis-ichthyosis-deafness syndrome
Keratitis-ichthyosis-deafness syndrome, atypical	Keratitis-ichthyosis-deafness syndrome
Keratitis-ichthyosis-deafness syndrome, autosomal dominant	Keratitis-ichthyosis-deafness syndrome
Keratitis-ichthyosis-deafness syndrome, autosomal dominant; Deafness, autosomal recessive 1A	Keratitis-ichthyosis-deafness syndrome
Keratitis-ichthyosis-deafness syndrome, autosomal dominant; Hystrix-like ichthyosis with deafness	Keratitis-ichthyosis-deafness syndrome
Keratoderma palmoplantar deafness	Keratoderma palmoplantar deafness
Keratoderma palmoplantar deafness; Deafness, autosomal dominant 3a	Keratoderma palmoplantar deafness
Late-onset deafness	NSHL
Long QT syndrome	Long QT syndrome
Long QT syndrome & atrial fibrillation	Long QT syndrome
Long QT syndrome 1	Long QT syndrome
Long QT syndrome 1, recessive; Long QT syndrome, LQT1 subtype	Long QT syndrome
Long QT syndrome 1; Acquired susceptibility to long QT syndrome 1; KCNQ1-related acquired long QT syndrome; Long QT syndrome, LQT1 subtype	Long QT syndrome
Long QT syndrome 1; Jervell and Lange-Nielsen syndrome	JLNS/Long QT syndrome
Long QT syndrome 1; KCNQ1-related Jervell and Lange-Nielsen syndrome; Long QT syndrome, LQT1 subtype	JLNS/Long QT syndrome
Long QT syndrome 1; Long QT syndrome 1/2, digenic; Long QT syndrome, LQT1 subtype	Long QT syndrome
Long QT syndrome 1; Long QT syndrome, LQT1 subtype	Long QT syndrome
Long QT syndrome, LQT1 subtype; Jervell and Lange-Nielsen syndrome; KCNQ1-related Jervell and Lange-Nielsen syndrome	JLNS/Long QT syndrome
Long QT syndrome, LQT1 subtype; Long QT syndrome 1	Long QT syndrome
Long QT syndrome, LQT1 subtype; Long QT syndrome 1, recessive; KCNQ1-related Jervell and Lange-Nielsen syndrome	JLNS/Long QT syndrome
Long QT syndrome, modifier of	Long QT syndrome modifier
Mixed hearing loss	Mixed hearing loss
Mondini deformity	Mondini malformation
Neurosensory deafness	NSHL
Noise-induced hearing loss, susceptibility to, association	NIHL susceptibility
Non-syndromic autosomal recessive deafness	NSHL

Table A-1: Continued.

Non-syndromic hearing loss	NSHL
Non-syndromic hearing loss, autosomal recessive	NSHL
Nonsyndromic deafness	NSHL
Nonsyndromic hearing loss	NSHL
Nonsyndromic hearing loss, autosomal recessive	NSHL
Optic atrophy & sensorineural hearing loss	Optic atrophy and SNHL
Optic atrophy, autosomal dominant, with hearing impairment	Optic atrophy and SNHL
Palmoplantar keratoderma with hearing loss	Palmoplantar keratoderma with hearing loss
Pendred syndrome	Pendred syndrome
Pendred's syndrome	Pendred syndrome
Pendred's syndrome; Hereditary hearing loss and deafness	Pendred syndrome
Peripheral neuropathy & hearing impairment	Peripheral neuropathy & hearing impairment
Phenotype modifier in Usher syndrome	Usher syndrome modifier
Phenotype modifier, association with	Usher syndrome modifier
Progressive hearing impairment	NSHL
Progressive hearing loss	NSHL
Progressive hearing loss, autosomal recessive	NSHL
Progressive hearing loss, nonsyndromic	NSHL
Progressive hearing loss, X-linked	XLNSHL
Retinitis pigmentosa & hearing loss	Usher syndrome
Retinitis pigmentosa 39; Retinitis pigmentosa; Retinitis pigmentosa-deafness syndrome	Usher syndrome
Retinitis pigmentosa 61	Retinitis Pigmentosa
Retinitis pigmentosa-deafness syndrome	Usher syndrome
Retinitis pigmentosa, autosomal recessive	Retinitis Pigmentosa
Retinitis pigmentosa, mild & deafness	Usher syndrome
Retinitis pigmentosa, nonsyndromic	Retinitis Pigmentosa
Retinitis pigmentosa; Retinitis pigmentosa-deafness syndrome	Usher syndrome
Sensorineural deafness	NSHL
Sensorineural deafness with hypertrophic cardiomyopathy	SNHL with HCM
Sensorineural deafness with palmoplantar lichen planus	SNHL with palmoplantar lichen planus
Sensorineural deafness, nonsyndromic	NSHL
Sensorineural hearing loss	NSHL

Table A-1: Continued.

Sensorineural hearing loss, nonsyndromic	NSHL
SeSAME syndrome	Seizures, sensorineural deafness, ataxia, mental retardation, and electrolyte imbalance; SeSAME syndrome
Usher syndrome	Usher syndrome
Usher syndrome 1	Usher syndrome
Usher syndrome 1b	Usher syndrome
Usher syndrome 1c	Usher syndrome
Usher syndrome 1d	Usher syndrome
Usher syndrome 1f	Usher syndrome
Usher syndrome 1g	Usher syndrome
Usher syndrome 1j	Usher syndrome
Usher syndrome 2	Usher syndrome
Usher syndrome 2a	Usher syndrome
Usher syndrome 3	Usher syndrome
Usher syndrome 3a	Usher syndrome
Usher syndrome, atypical	Usher syndrome
Usher syndrome, type 1B	Usher syndrome
Usher syndrome, type 1B; AllHighlyPenetrant	Usher syndrome
Usher syndrome, type 1B; Retinitis pigmentosa-deafness syndrome	Usher syndrome
Usher syndrome, type 1C	Usher syndrome
Usher syndrome, type 1C; AllHighlyPenetrant	Usher syndrome
Usher syndrome, type 1D; Hereditary hearing loss and deafness; Retinitis pigmentosa-deafness syndrome	Usher syndrome
Usher syndrome, type 1F	Usher syndrome
Usher syndrome, type 1F; Hereditary hearing loss and deafness; Retinitis pigmentosa-deafness syndrome; Usher syndrome, type 1G	Usher syndrome
Usher syndrome, type 1G	Usher syndrome
Usher syndrome, type 1J	Usher syndrome
Usher syndrome, type 2A	Usher syndrome
Usher syndrome, type 2A; Retinitis pigmentosa 39; Retinitis pigmentosa-deafness syndrome	Usher syndrome
Usher syndrome, type 2A; Retinitis pigmentosa-deafness syndrome	Usher syndrome
Usher syndrome, type 2C	Usher syndrome
Usher syndrome, type 2D	Usher syndrome

Table A-1: Continued.

Usher syndrome, type 3	Usher syndrome
Usher syndrome, type 3; Retinitis pigmentosa-deafness syndrome	Usher syndrome
USHER SYNDROME, TYPE ID/F, DIGENIC; AllHighlyPenetrant	Usher syndrome
X-chromosomal hearing loss	XLNSHL
X-linked deafness	XLNSHL
X-linked mixed deafness	XLNSHL

Table A-2: GJB2 variants selected from HGMD to check for discrepancies between the clinical significance provided in HGMD and the respective references provided in HGMD.

Nucleotide change	Amino acid change	Reference(s)
c.584T>C	p.M195T	(Lee *et al.*, 2009)
c.247T>A	p.F83I	(Lee *et al.*, 2009)
c.412A>G	p.S138G	(Picciotti *et al.*, 2009)
c.263C>G	p.A88G	(Alemanno *et al.*, 2009)
c.514T>A	p.W172R	(Mani *et al.*, 2009)
c.98T>C	p.I33T	(Mani *et al.*, 2009)
c.548C>T	p.S183F	(de Zwart-Storm *et al.*, 2008)
c.203A>G	p.Y68C	(Roux *et al.*, 2004; Tsukada *et al.*, 2010)
c.124G>A	p.E42K	(Wu *et al.*, 2008)
c.53C>T	p.T18I	(Guo *et al.*, 2008)
c.149A>C	p.D50A	(Mhaske *et al.*, 2013)
c.37G>A	p.V13M	(Putcha *et al.*, 2007)
c.91T>A	p.F31I	(Putcha *et al.*, 2007)
c.653G>A	p.C218Y	(Putcha *et al.*, 2007)
c.191G>A	p.C64Y	(Putcha *et al.*, 2007)
c.139G>C	p.E47Q	(Putcha *et al.*, 2007)
c.650A>G	p.Y217C	(Putcha *et al.*, 2007)
c.677T>G	p.V226G	(Putcha *et al.*, 2007)
c.473A>G	p.Y158C	(Putcha *et al.*, 2007)
c.499G>A	p.V167M	(Putcha *et al.*, 2007)
c.389G>A	p.G130D	(Putcha *et al.*, 2007)
c.241C>G	p.L81V	(Putcha *et al.*, 2007)
c.358G>A	p.E120K	(Putcha *et al.*, 2007)
c.167T>C	p.L56P	(Putcha *et al.*, 2007)
c.17T>C	p.L6P	(Putcha *et al.*, 2007)
c.227T>C	p.L76P	(Batissoco *et al.*, 2009)
c.109G>C	p.V37L	(Putcha *et al.*, 2007)
c.250G>T	p.V84L	(Bruzzone *et al.*, 2003)
c.428G>T	p.R143L	(Putcha *et al.*, 2007)

Table A-2: Continued.

Nucleotide change	Amino acid change	Reference(s)
c.209C>T	p.P70L	(Putcha *et al.*, 2007)
c.488T>C	p.M163T	(Putcha *et al.*, 2007)
c.278T>C	p.M93T	(Putcha *et al.*, 2007)
c.557C>A	p.T186K	(Putcha *et al.*, 2007)
c.563A>G	p.K188R	(Putcha *et al.*, 2007)
c.161A>G	p.N54S	(Putcha *et al.*, 2007)
c.172C>G	p.P58A	(Primignani *et al.*, 2007)
c.314A>G	p.K105R	(de Oliveira *et al.*, 2007)
c.160A>C	p.N54H	(Bazazzadegan *et al.*, 2011)
c.61G>A	p.G21R	(Rabionet *et al.*, 2006)
c.188T>C	p.V63A	(Tang *et al.*, 2006)
c.40A>G	p.N14D	(Haack *et al.*, 2006)
c.107T>C	p.L36P	(Propst, Papsin, *et al.*, 2006)
c.380G>T	p.R127L	(Tang *et al.*, 2006)
c.475G>T	p.D159Y	(Propst, Stockley, *et al.*, 2006)
c.257C>T	p.T86M	(Tang *et al.*, 2006)
c.40A>T	p.N14Y	(Mhaske *et al.*, 2013)
c.102G>A	p.M34I	(Snoeckx *et al.*, 2005)
c.154G>C	p.V52L	(Snoeckx *et al.*, 2005)
c.299A>C	p.H100P	(Snoeckx *et al.*, 2005)
c.394C>G	p.L132V	(Snoeckx *et al.*, 2005)
c.622A>C	p.T208P	(Snoeckx *et al.*, 2005)
c.452T>G	p.M151R	(Snoeckx *et al.*, 2005)
c.419T>G	p.I140S	(Snoeckx *et al.*, 2005)
c.413G>A	p.S138N	(Snoeckx *et al.*, 2005)
c.389G>T	p.G130V	(Iossa *et al.*, 2009)
c.175G>C	p.G59R	(Leonard *et al.*, 2005)
c.164C>A	p.T55N	(Tekin *et al.*, 2005)
c.617A>C	p.N206T	(Wattanasirichaigoon *et al.*, 2004)
c.119C>T	p.A40V	(Mhaske *et al.*, 2013)
c.331A>G	p.I111V	(Azaiez *et al.*, 2004)
c.506G>A	p.C169Y	(Azaiez *et al.*, 2004)

Table A-2: Continued.

Nucleotide change	Amino acid change	Reference(s)
c.110T>C	p.V37A	(Azaiez *et al.*, 2004)
c.42C>G	p.N14K	(Lazic *et al.*, 2008)
c.176G>T	p.G59V	(Palmada *et al.*, 2006)
c.50C>A	p.S17Y	(Tóth *et al.*, 2004)
c.475G>A	p.D159N	(Snoeckx *et al.*, 2005)
c.268C>G	p.L90V	(Snoeckx *et al.*, 2005)
c.187G>A	p.V63M	(Snoeckx *et al.*, 2005)
c.249C>G	p.F83L	(Bruzzone *et al.*, 2003; Roux *et al.*, 2004)
c.251T>C	p.V84A	(Ferraris *et al.*, 2002)

REFERENCES

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

Alemanno,M.S. *et al.* (2009) A novel missense mutation in the Connexin 26 gene associated with autosomal recessive nonsyndromic sensorineural hearing loss in a consanguineous Tunisian family. *Int. J. Pediatr. Otorhinolaryngol.*, **73**, 127–131.

Azaiez,H. *et al.* (2004) GJB2: the spectrum of deafness-causing allele variants and their phenotype. *Hum. Mutat.*, **24**, 305–311.

Batissoco,A.C. *et al.* (2009) A novel missense mutation p.L76P in the GJB2 gene causing nonsyndromic recessive deafness in a Brazilian family. *Braz. J. Med. Biol. Res. Rev. Bras. Pesqui. Médicas E Biológicas Soc. Bras. Biofísica Al*, **42**, 168–171.

Bazazzadegan,N. *et al.* (2011) Two Iranian families with a novel mutation in GJB2 causing autosomal dominant nonsyndromic hearing loss. *Am. J. Med. Genet. A.*, **155A**, 1202–1211.

Bruzzone,R. *et al.* (2003) Loss-of-function and residual channel activity of connexin26 mutations associated with non-syndromic deafness. *FEBS Lett.*, **533**, 79–88.

Carter,H. *et al.* (2013) Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics*, **14**, S3.

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

Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

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**.

Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, USA.

Ferraris,A. *et al.* (2002) Pyrosequencing for detection of mutations in the connexin 26 (GJB2) and mitochondrial 12S RNA (MTRNR1) genes associated with hereditary hearing loss. *Hum. Mutat.*, **20**, 312–320.

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

Garber,M. *et al.* (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**, i54–i62.

Guo,Y.-F. *et al.* (2008) GJB2, SLC26A4 and mitochondrial DNA A1555G mutations in prelingual deafness in Northern Chinese subjects. *Acta Otolaryngol. (Stockh.)*, **128**, 297–303.

Haack,B. *et al.* (2006) Deficient membrane integration of the novel p.N14D-GJB2 mutant associated with non-syndromic hearing impairment. *Hum. Mutat.*, **27**, 1158–1159.

Iossa,S. *et al.* (2009) New evidence for the correlation of the p.G130V mutation in the GJB2 gene and syndromic hearing loss with palmoplantar keratoderma. *Am. J. Med. Genet. A.*, **149A**, 685–688.

Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.

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.

Lazic,T. *et al.* (2008) A report of GJB2 (N14K) Connexin 26 mutation in two patients--a new subtype of KID syndrome? *Pediatr. Dermatol.*, **25**, 535–540.

Lee,K.H. *et al.* (2009) Audiologic and temporal bone imaging findings in patients with sensorineural hearing loss and GJB2 mutations. *The Laryngoscope*, **119**, 554–558.

Leonard,N.J. *et al.* (2005) Sensorineural hearing loss, striate palmoplantar hyperkeratosis, and knuckle pads in a patient with a novel connexin 26 (GJB2) mutation. *J. Med. Genet.*, **42**, e2.

Li,H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

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.

Liu,X. *et al.* (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, **32**, 894–899.

Mani,R.S. *et al.* (2009) Functional consequences of novel connexin 26 mutations associated with hereditary hearing loss. *Eur. J. Hum. Genet. EJHG*, **17**, 502–509.

Mhaske,P.V. *et al.* (2013) The human Cx26-D50A and Cx26-A88V mutations causing keratitis-ichthyosis-deafness syndrome display increased hemichannel activity. *Am. J. Physiol. Cell Physiol.*, **304**, C1150–1158.

De Oliveira,C.A. *et al.* (2007) Molecular genetics study of deafness in Brazil: 8-year experience. *Am. J. Med. Genet. A.*, **143A**, 1574–1579.

Palmada,M. *et al.* (2006) Loss of function mutations of the GJB2 gene detected in patients with DFNB1-associated hearing impairment. *Neurobiol. Dis.*, **22**, 112–118.

Picciotti,P.M. *et al.* (2009) Correlation between GJB2 mutations and audiological deficits: personal experience. *Eur. Arch. Oto-Rhino-Laryngol. Off. J. Eur. Fed. Oto-Rhino-Laryngol. Soc. EUFOS Affil. Ger. Soc. Oto-Rhino-Laryngol. - Head Neck Surg.*, **266**, 489–494.

Primignani,P. *et al.* (2007) A new de novo missense mutation in connexin 26 in a sporadic case of nonsyndromic deafness. *Laryngoscope*, **117**, 821–824.

Propst,E.J., Papsin,B.C., *et al.* (2006) Auditory responses in cochlear implant users with and without GJB2 deafness. *Laryngoscope*, **116**, 317–327.

Propst,E.J., Stockley,T.L., *et al.* (2006) Ethnicity and mutations in GJB2 (connexin 26) and GJB6 (connexin 30) in a multi-cultural Canadian paediatric Cochlear Implant Program. *Int. J. Pediatr. Otorhinolaryngol.*, **70**, 435–444.

Putcha,G.V. *et al.* (2007) A multicenter study of the frequency and distribution of GJB2 and GJB6 mutations in a large North American cohort. *Genet. Med. Off. J. Am. Coll. Med. Genet.*, **9**, 413–426.

Rabionet,R. *et al.* (2006) A novel G21R mutation of the GJB2 gene causes autosomal dominant non-syndromic congenital deafness in a Cuban family. *Genet. Mol. Biol.*, **29**, 443–445.

Reva,B. *et al.* (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118–e118.

Roux,A.-F. *et al.* (2004) Molecular epidemiology of DFNB1 deafness in France. *BMC Med. Genet.*, **5**, 5.

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.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Shihab,H.A. *et al.* (2013) Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Hum. Mutat.*, **34**, 57–65.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

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'06. Springer-Verlag, Berlin, Heidelberg, pp. 190–205.

Snoeckx,R.L. *et al.* (2005) GJB2 Mutations and Degree of Hearing Loss: A Multicenter Study. *Am. J. Hum. Genet.*, **77**, 945–957.

Stenson,P.D. *et al.* (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9.

Tang,H.-Y. *et al.* (2006) DNA sequence analysis of GJB2, encoding connexin 26: observations from a population of hearing impaired cases and variable carrier rates, complex genotypes, and ethnic stratification of alleles among controls. *Am. J. Med. Genet. A.*, **140**, 2401–2415.

Tekin,M. *et al.* (2005) Evidence for single origins of 35delG and delE120 mutations in the GJB2 gene in Anatolia. *Clin. Genet.*, **67**, 31–37.

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

Tóth,T. *et al.* (2004) GJB2 mutations in patients with non-syndromic hearing loss from Northeastern Hungary. *Hum. Mutat.*, **23**, 631–632.

Tsukada,K. *et al.* (2010) A large cohort study of GJB2 mutations in Japanese hearing loss patients. *Clin. Genet.*, **78**, 464–470.

Wattanasirichaigoon,D. *et al.* (2004) High prevalence of V37I genetic variant in the connexin-26 (GJB2) gene among non-syndromic hearing-impaired and control Thai individuals. *Clin. Genet.*, **66**, 452–460.

Wu,C.-C. *et al.* (2008) Prospective mutation screening of three common deafness genes in a large Taiwanese Cohort with idiopathic bilateral sensorineural hearing impairment reveals a difference in the results between families from hospitals and those from rehabilitation facilities. *Audiol. Neurootol.*, **13**, 172–181.

De Zwart-Storm,E.A. *et al.* (2008) A novel missense mutation in the second extracellular domain of GJB2, p.Ser183Phe, causes a syndrome of focal palmoplantar keratoderma with deafness. *Am. J. Pathol.*, **173**, 1113–1119.