AACDS: A database for personal genome interpretation


Thanawadee Preeprem✉, **Greg Gibson**
Georgia Institute of Technology, Atlanta, United States

Received 11 August 2014; **Accepted** 5 September 2014; **Published** 10 October 2014

Preeprem T and Gibson G. (2014) *EMBnet.journal* **20**, e780. http://dx.doi.org/10.14806/ej.20.0.780

Competing Interests: none

Abstract
Incorporation of diverse data sources adds value to genomic studies, especially for annotation and categorisation of personal genome variants. The database for Association-Adjusted Consensus Deleterious Scheme (AACDS) and its Web application deliver a novel approach to assess genetic variations based on their putative functionality. The database is built upon integrated knowledge of variant data, with the aim of relating clinical phenotypes to predictions of variant deleteriousness. The simple but inter-related queries classify each variant into an 8-level category. The categories can be ranked, enabling straightforward interpretation of relative likelihood of functionality. The ranking thus facilitates improved efficiency in prioritising further detailed evaluation of key variants within a personal genome. The AACDS database covers more than 68 million mis-sense variants in approximately 18,000 human genes. Given a list of genetic variants, the retrieval of the AACDS category, along with known clinical data can be performed through an intuitive search platform.
Availability: The AACDS Web application is publicly available at http://cig.gatech.edu/tools.

Introduction
Non-synonymous Single Nucleotide Polymorphism (nsSNP) is one of the most common forms of genomic variability. About 60% of known disease-causing mutations are nsSNPs (Cooper *et al.*, 2010). One of the major goals for personal genomics is to identify a subset of variants that have the potential to influence an individual's health. Each individual genome is estimated to contain roughly ten thousand nsSNPs (Kim *et al.*, 2009; Ng *et al.*, 2008; Patel, *et al.*, 2013). The assessment of deleteriousness for SNPs is commonly performed on a per variant basis, by using many available computational tools that typically classify each SNP into two groups: benign and damaging. Although many prediction programmes have been proven to have acceptable accuracy, mostly in the range of 70-80% (Gonzalez-Perez and Lopez-Bigas, 2011), it is deemed an advantage to incorporate more data into the assessment (Ng and Henikoff, 2006).

In our recent study on interpretation of personal genome data (Preeprem and Gibson, 2013), we developed the "Association-Adjusted Consensus Deleterious Scheme" (AACDS) to facilitate variant prioritisation of personal genome studies. AACDS is constructed from the combination of existing databases that implicate the variant with disease or phenotype, and traditional sequence-based predictions. It classifies variants according to an 8-level category. Not only does AACDS incorporate the clinical or phenotypic annotations of the genomic variants in an individual, it also narrows down the variants to a subset that is appropriate for further follow-up experiments and validation with respect to individualised health profiles.

To promote the utility of our variant classification schema, AACDS, we have implemented the assessments into a database-driven Web application that allows users to search the AACDS categories and relevant information for user-defined variants. The AACDS website aims to provide a user-friendly platform for anyone interested in personal genome interpretation. The database schema was designed to cover the annotated list of functional variants (31,092 disease-associated amino acid variants in 3,363 genes), 4,225 pairs of gene-disease associations, 5,113 pairs

of gene-trait associations, and all possible coding genomic variants in 18,349 human genes (*i.e.*, 68,165,196 nsSNPs). Therefore, our newly developed database-driven Web application for AACDS can serve as a tool to generate the best estimate of clinical significance of each variant from the large and growing accumulation of personal genome data. In addition to identifying causal variants or variants in disease- or trait-associated genes from a list of genomic variability, the application also allows users to carry out further functional analyses of all SNPs in any gene of interest.

Although many tools and databases exist for the purpose of variant prioritisation and/or personal genome interpretation, we are not aware of any tool with similar features to ours, especially in the categorisation of genomic variants. Our AACDS tool allows SNP evaluations to be performed simultaneously on the basis of deleterious predictions, direct connections between variants to diseases, and associated traits and diseases to the genes. The tool assigns an AACDS class to each individual SNP; it also reports the overall AACDS statistics for a given genome. The classification and ranking of SNPs is particularly significant and original, as it assists effortless interpretations of whole-genome SNP searches. The results facilitate the identification of high-impact variants within a genome in an effective and efficient manner.

Compared to aggregative variant association methods such as in VAAST 2.0 (Hu *et al.*, 2013), our tool does not require that users have prior knowledge of various additional genomic attributes to perform searches and interpret the results. VAAST requires not only target and background genome data-sets, but also user-defined sets of genes and prior knowledge of genetic parameters (inheritance, penetrance, locus heterogeneity, allele frequency, *etc.*) in order to search for causal SNPs or genes. The search pipeline is neither designed for evaluation of all genomic variants, nor as a simple look-up utility.

Two recent genome analysis tools, eXtasy (Sifrim *et al.*, 2013) and Phen-Gen (Javed *et al.*, 2014), introduce a new phase of genome interpretation, in which the tools link genome variants to a specific phenotype. Although both tools have great potential for guiding diagnostics of rare disorders through the identification of phenotype-specific causal variants, the evaluations are performed on a per disease basis. Most personal genome variants are likely to be neutral and contain a minimal number of annotated disease SNPs (Preeprem and Gibson, 2013; Xue *et al.*, 2012); the individuals are healthy and unlikely to have noticeable clinical phenotypes (Patel *et al.*, 2013). These limitations represent a significant challenge for personal genome variant annotation for sub-clinical phenotypes, in whose interpretation AACDS is designed to help.

Implementation

The AACDS website serves as an interface for queries of the AACDS database, which is built to categorise nsSNPs into an 8-level class, based on their consensus predicted deleteriousness and the evidence of disease or complex trait associations with their genes. The database includes 68,165,196 nsSNPs that can be found in a human genome. The website allows users to retrieve the AACDS classification and relevant information about variants in genes of interest.

Data sources

To facilitate the variant mapping of various data types (chromosome coordinates, gene names, protein names), we chose UniProtKB (UniProt Consortium, 2012) as the core database. UniProtKB accession numbers provide unique identifiers for gene products, allowing direct look-up of the disease-association data from the selected SNP databases: MSV3d (Luu *et al.*, 2012) and SwissVar (Mottaz *et al.*, 2010). A list of 20,277 reviewed human proteins (representing the gene products of 19,700 genes) was compiled from UniProtKB (2012_06 release, accessed 1 November 2013).

Next, we used dbNSFP v2.1 (Liu *et al.*, 2011) (released 3 October 2013) to extract all possible SNP locations within each gene. The database provides translations of nucleotide variants into alternate amino acids, which we indexed with respect to the corresponding proteins. All SNP functional predictions (benign vs. damaging) were retrieved from the pre-computed scores for six sequence-based deleterious predictors available from dbNSFP. To resolve discrepancies among prediction algorithms, we assigned levels of deleteriousness using the consensus prediction. A variant is regarded as "deleterious" if $\geq$ 3/6 predictors reported the variant as "deleterious", and as "non-deleterious" if the predictions

suggest otherwise. Later, the initial set of SNPs was filtered such that only variants located in known genes were retained (68,165,196 nsSNP locations in 18,349 genes).

Gene-trait associations were retrieved from the NHGRI Genome-Wide Association Studies (GWAS) catalogue (Hindorff, *et al.*), available from dbNSFP v2.1. Additional information provided at the AACDS website includes essential information about each variant: *i.e.*, dbSNP reference SNP ID number (db138 release, downloaded from the NCBI's FTP site at ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/BED, accessed 16 January 2014), gene name and protein name from UniProtKB, and population-specific minor allele frequencies (retrieved from dbNSFP).

Database construction

AACDS was designed as a relational database on a MySQL server. The data relationships are presented in Figure 1.

In-house Perl scripts were used to extract variant information from the aforementioned data sources. Our original paper describes the AACDS as an 8-level category (variant categories 1, 2A, 2B, 3A, 3B, 4, 5, and 6) (Preeprem and Gibson, 2013). However, many SNPs cannot be exclusively defined into one class; therefore, a maximum of 12 classes are reported in this implementation to represent all distinct conditions possible when joining multiple assigned AACDS categories together (Table 1).

The list of disease associations was collected from SwissVar (accessed 1 November 2013) and MSV3d (released 29 July 2012) databases. We did not attempt to standardise the minor differences of clinical terms provided by the two data sources. Similar association records for a particular SNP or a gene from both SNP databases were dealt with by reporting only the most detailed record. Some SNPs have ambiguous clinical annotations; for example, when one of the two databases documents a SNP as a disease-associated variant, but the other suggests it is a polymorphism or has missing data, the intuition we followed was to regard the variant to have

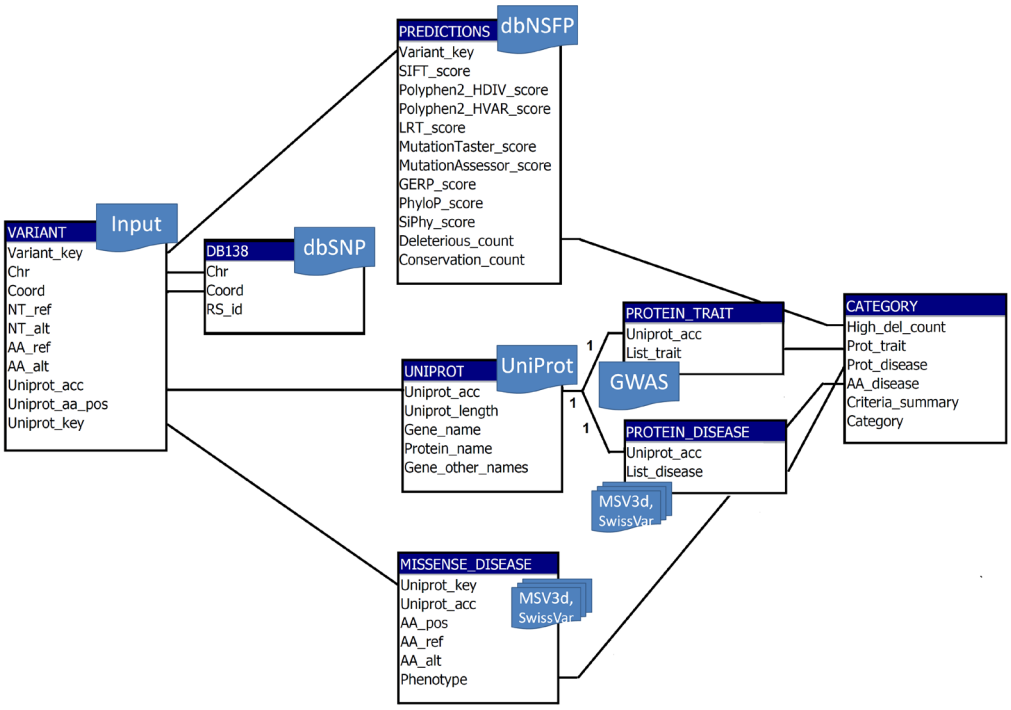

Figure 1. AACDS database schema. The database constructs its data relationships from several sources. The AACDS category for each variant is assigned according to whether the variant has a high deleterious count, whether its gene has GWAS-documented gene-trait/gene-disease association, its gene is database-documented to have disease association, or the variant is documented as a disease-causal variant.

clinical associations. In total, 31,092 instances of variant-disease associations and 4,225 pairs of gene-disease associations were included in our database. The number of genes whose gene-trait associations were identified from GWAS is 5,113.

To ensure that search results are returned quickly, we performed the computation of AACDS for all variants, and utilised the assigned categories as the pre-computed variant classification during Web searching. The online service of the AACDS database was implemented in PHP, MySQL, JavaScript and Apache. The AACDS

Table 1. Descriptions of the 12 combined AACDS classes. Column descriptions for features of nsSNPs are (i) disease-causing, if MSV3d and/or SwissVar indicate the variant is disease-causal; (ii) predicted deleterious, if ≥ 3/6 programmes predict the variant to be deleterious; (iii) in disease gene, if MSV3d and/or SwissVar indicate the gene has disease associations; (iv) in GWAS-documented trait/disease gene, if GWAS indicates the gene has trait/disease associations.

AACDS classes	Features of nsSNPs				Descriptions of nsSNPs
	(i) Disease-causing	(ii) Predicted deleterious	(iii) In disease gene	(iv) In GWAS-documented trait/disease gene	
1	X				disease-causing (but not located in gene with disease- or trait-associations nor predicted as deleterious by most programmes)
1, 2B	X	X	X		disease-causing, predicted as deleterious by most programmes, located in gene with disease-associations (but not GWAS-documented)
1, 2B, 3B	X	X	X	X	disease-causing, predicted as deleterious by most programmes, located in gene with disease and trait-associations
1, 5	X		X	X	disease-causing, located in gene with disease- and trait-associations (but most programmes predicted it to be benign)
2A			X		located in gene with database-documented disease-associations (but no other implications)
2B		X	X		predicted deleterious by most programmes, located in gene with database-documented disease-associations (but not a causal variant)
2B, 3B		X	X	X	predicted deleterious by most programmes, located in gene with disease and trait-associations (but not a causal variant)
3A				X	located in gene with GWAS-documented trait/disease associations (but no other implications)
3B		X		X	predicted deleterious by most programmes, located in gene with GWAS-documented trait/disease associations (but not a causal variant)
4		X			predicted deleterious by most programmes (but no other implications)
5			X	X	located in gene with disease and trait-associations (but not a causal variant nor predicted deleterious)
6					no clinical implications

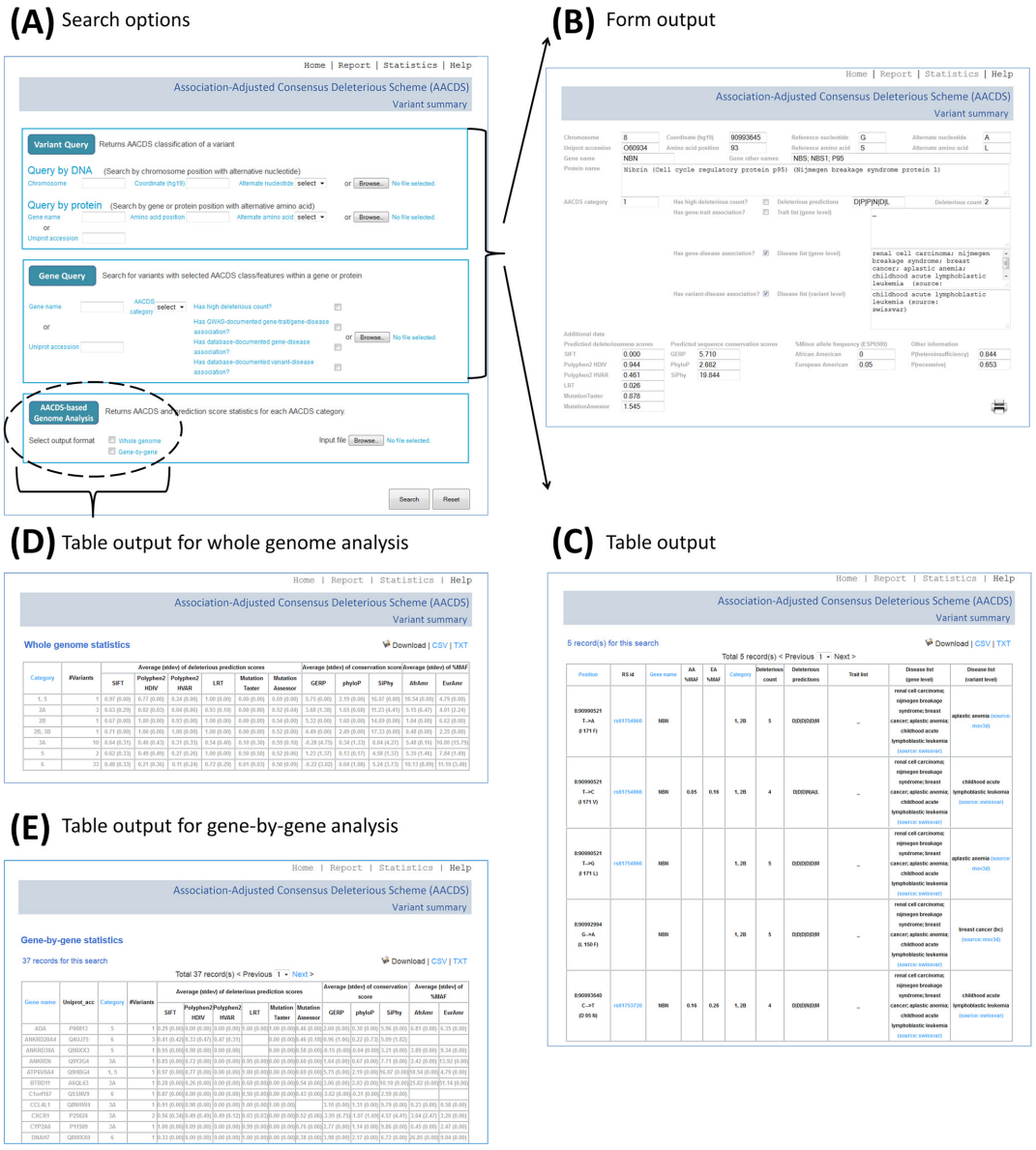

Figure 2. Overview of the AACDS Web interface. (A) The three query options: Variant query, Gene query, and AACDS-based genome analysis. (B-E) Example outputs in form and tabular formats. The form output (B) reports the AACDS category of a variant and its relevant information, along with any additional variant data. Included in the tabular output (C) are direct links to dbSNP and to the original sources of clinical data. The outputs from AACDS-based genome analysis (D-E) present numerical statistics of nsSNPs based on the assigned AACDS classes.

website can be accessed at http://cig.gatech.edu/tools. All standard browsers are supported.

Utility

Our AACDS Web application allows users to retrieve AACDS classifications and relevant information about variants or genes of interest. Figure 2A illustrates the three major components of the website: (1) Variant query, (2) Gene query, and (3) AACDS-based genome analysis. Users can search the AACDS database via single-query or batch mode. Batch mode permits practical analysis of personal genome data, as users can upload lists of variants of unlimited size and retrieve the results in plain-text formats for external use.

Table 2. File formats for batch queries. The following analyses accept a batch search if users provide a .txt file (tab delimited) with a specified format.

Queries	Query options	File formats	Column descriptions
Variant query	By DNA	Chr:10 26781257 T A Chr:10 26781257 T C Chr:10 26781257 T G	1 = chromosome number 2 = hg19 coordinate 3 = reference nucleotide 4 = alternative nucleotide
	By protein (providing gene or protein names)	Gene:AACS 8 G S Gene:GOT1 413 Q H Gene:NT5C2 515 K Q Or Uniprot:Q8IZY2 2000 N K Uniprot:Q86UK0 2000 T A Uniprot:O95477 2000 L R	1 = gene name or UniProtKB accession number 2 = amino acid position (UniProtKB numbering) 3 = reference amino acid 4 = alternative amino acid
Gene query	Providing gene or protein names	Gene:HSD3B2 1 Gene:ABCA12 1 Gene:SH3BP2 1 Or Uniprot:Q86V21 4 Uniprot:P01011 2B Uniprot:Q9NY61 4	1 = gene name, or UniProtKB accession number 2 = AACDS category (1, 2A, 2B, 3A, 3B, 4, 5, or 6)
AACDS-based genome analysis	-	Chr:10 26781257 T A Chr:10 26781257 T C Chr:10 26781257 T G	1 = chromosome number 2 = hg19 coordinate 3 = reference nucleotide 4 = alternative nucleotide

For a single-entry query, users can search for the AACDS classification of their variant of interest by providing some search parameters: for query by DNA, chromosome number, hg19 coordinate and alternative nucleotide; for query by protein, gene name or UniProtKB accession number, amino acid position, and alternative amino acid. The website outputs a variant summary page, which reports the AACDS category of the variant and its relevant information, along with any additional variant data (Figure 2B).

Users can also retrieve lists of gene variants whose characteristics match their interests. If a particular AACDS class is specified, the website returns all nsSNPs that belong to that category. If any of the four features (Table 1, Figure 2A) are specified, a list of variants whose characteristics are compatible with the search features is returned. When more than one variant meets the search criteria, a form (Figure 2B) and summary table (Figure 2C) are returned. The table provides a short description (11 attributes) of the variants; users can also download the complete table (37 attributes) through the "download" button.

We also provide the overall statistics for a set of nsSNPs found in an individual's genome via the AACDS-based genome analysis option. Users can perform the analysis on two levels: whole genome statistics and gene-by-gene statistics – Figures 2D and 2E show example outputs from the two analysis types, respectively. In either case, the schema classifies nsSNPs into several groups, based on the assigned AACDS classes. The results can be ranked by gene names or by AACDS groups. In addition to the number of variants within each AACDS class, the tabular output also presents the average (and the standard deviation) for all six deleterious scores, three conservation scores, and two population-specific minor allele frequencies.

For each of the above analyses, a batch search is possible. The required information for input file formats is described in Table2.

Discussion

The integration of both sequence-based deleterious prediction and clinical-association data in our AACDS algorithm provides a novel approach to integrative variant classification for personal genomes. Manual inspection of a variant for both predicted deleteriousness and phenotypic association is possible, but certainly not practical

for analysing large genome data. For this reason, the implementation of a database-driven Web application is considered an important tool for promoting the utility of the AACDS. We believe that with the scope of our database coverage, both in terms of genomic variations and phenotypic data, this application will help to bring a comprehensive framework of personal genome interpretation to a more practical level.

The current implementation does not have an automatic online update feature, but we will regularly check for new releases of our selected external databases so that it offers AACDS classes for the most complete set of SNPs in a human genome. Further improvements may include subsequent addition of variants in the remaining genes once their curated protein sequences are available, the inclusion of clinical and trait associations from other data sources, and the implementation of an automatic online update with the selected data sources.

Key Points

- Association-Adjusted Consensus Deleterious Scheme (AACDS) is an integrative approach for interpreting genomic variations, using variant deleteriousness predictors and publicly available genomics data.
- AACDS is specifically designed for personal genome analysis (variants likely to be neutral).
- AACDS database covers all missense variants (induce amino acid changes) of over 18,000 human genes.
- AACDS Web application enables the evaluations of variants on a per variant, per gene, and per genome basis.
- AACDS facilities the identification and prioritisation of significant variants.

References

Cooper DN, Chen JM, Ball EV, Howells K, Mort M, *et al.* (2010) Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Human Mutation* **31**, 631-655. http://dx.doi.org/10.1002/humu.21260

Gonzalez-Perez A and Lopez-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. American *Journal of Human Genetics* **88**, 440-449. http://dx.doi.org/10.1016/j.ajhg.2011.03.004

Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, *et al.* A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies.

Hu H, Huff CD, Moore B, Flygare S, Reese MG, *et al.* (2013) VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology* **37**, 622-634. http://dx.doi.org/10.1002/gepi.21743

Javed A, Agrawal S and Ng PC (2014) Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nature Methods* **11**, 935–937. http://dx.doi.org/10.1038/nmeth.3046

Kim JI, Ju YS, Park H, Kim S, Lee S, *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual, *Nature* **460**, 1011-1015. http://dx.doi.org/10.1038/nature08211

Liu X, Jian X and Boerwinkle E (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions, *Human Mutation* **32**, 894-899. http://dx.doi.org/10.1002/humu.21517

Luu TD, Rusu AM, Walter V, Ripp R, Moulinier L, *et al.* (2012) MSV3d: database of human MisSense Variants mapped to 3D protein structure, *Database* **2012**, bas018. http://dx.doi.org/10.1093/database/bas018

Mottaz A, David FP, Veuthey AL and Yip YL (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar, *Bioinformatics* **26**, 851-852. http://dx.doi.org/10.1093/bioinformatics/btq028

Ng PC and Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function, *Annual Review of Genomics and Human Genetics* **7**, 61-80. http://dx.doi.org/10.1146/annurev.genom.7.080505.115630

Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, *et al.* (2008) Genetic variation in an individual human exome, *PLoS Genetics* **4**, e1000160. http://dx.doi.org/10.1371/journal.pgen.1000160

Patel CJ, Sivadas A, Tabassum R, Preeprem T, Zhao J, *et al.* (2013) Whole genome sequencing in support of wellness and health maintenance, *Genome Medicine* **5**, 58. http://dx.doi.org/10.1186/gm462

Preeprem T and Gibson G (2013) An association-adjusted consensus deleterious scheme to classify homozygous Mis-sense mutations for personal genome interpretation, *BioData Mining*, **6**, 24. http://dx.doi.org/10.1186/1756-0381-6-24

Sifrim A, Popovic D, Tranchevent LC, Ardeshirdavani A, Sakai R, *et al.* (2013) eXtasy: variant prioritization by genomic data fusion, *Nature Methods*, **10**, 1083-1084. http://dx.doi.org/10.1038/nmeth.2656

UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acids Research*, **40**, D71-75. http://dx.doi.org/10.1093/nar/gkr981

Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, *et al.* (2012) Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing, *American Journal of Human Genetics* **91**, 1022-1032. http://dx.doi.org/10.1016/j.ajhg.2012.10.015

Acknowledgements

This work was supported by start-up funds from the Georgia Tech Research Foundation to GG, and TP was supported by the School of Biology at Georgia Tech.