EMBnet.journal 20

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

e780

e780

Research Papers

EMBnet.journal 20

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 evalu-

ations 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

EMBnet.journal 20

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 <u>ftp://ftp.ncbi.nih.gov/</u> <u>snp/organisms/human_9606/BED</u>, 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 diseaseassociated 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.

EMBnet.journal 20

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

To ensure that search results are returned 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 \geq 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.

		Features	of nsSNPs		
AACDS classes	(i) Disease- causing	(ii) Predicted deleterious	(iii) In disease gene	(iv) In GWAS- documented trait/disease gene	Descriptions of nsSNPs
1	х				disease-causing (but not located in gene with disease- or trait-associations nor predicted as deleterious by most programmes)
1, 2B	Х	Х	Х		disease-causing, predicted as deleterious by most programmes, located in gene with disease-associations (but not GWAS-docu- mented)
1, 2B, 3B	Х	Х	Х	Х	disease-causing, predicted as deleterious by most programmes, located in gene with dis- ease and trait-associations
1,5	х		Х	Х	disease-causing, located in gene with dis- ease- and trait-associations (but most pro- grammes predicted it to be benign)
2A			Х		located in gene with database-documented disease-associations (but no other implica- tions)
2B		х	х		predicted deleterious by most programmes, located in gene with database-documented disease-associations (but not a causal vari- ant)
2B, 3B		Х	Х	Х	predicted deleterious by most programmes, located in gene with disease and trait-associ- ations (but not a causal variant)
ЗA				Х	located in gene with GWAS-documented trait/disease associations (but no other implications)
ЗB		х		Х	predicted deleterious by most programmes, located in gene with GWAS-documented trait/disease associations (but not a causal variant)
4		Х			predicted deleterious by most programmes (but no other implications)
5			х	х	located in gene with disease and trait-associ- ations (but not a causal variant nor predicted deleterious)
6					no clinical implications

EMBnet.journal 20

Research Papers

(B) Form output

(C) Table output



		nome	1 merore e	
	Ass	ociation-Adjusted Consensus	Deleterious	Scheme (AACD
				Variant summar
Variant Query Re	turns AACDS classificat	ion of a variant		
Query by DNA (s	Search by chromosome	position with alternative nucleotide)		
Chromosome	Coordinate (hg19)	Alternate nucleotide select •	or Browse	No file selected.
Query by protein	(Search by gene or prot	ein position with alternative amino acid)		
Jene name	Amino acid position	Alternate amino acid select 💌	or Browse.	No file selected.
Jriprot accession				
Gene Query S	earch for variants with se	lected AACDS classifeatures within a gene	or protein	
Gene Query S	earch for variants with se	lected AACDS classifeatures within a gene	or protein	
Gene Query St Gene name Of	AACDS category select •	Hected AACDS classificatures within a gene Has high deletarious count? Has GWMS-documented gene-traitigene-disease securities?	or protein	
Gene Query Se Jene name or	ACOS calepony select •	lected AACDS classificatures within a gene Has high deleterious count? Has DMA-documented gene-haligene-disease associator? Has addaba-documented gene-disease	or protein] No file selected.
Gene Query Si Jene name or Jriprot accession	earch for variants with se ACDS select •	Netted AACDS classificatures within a gene Has high deletinus court? Has 00%8-documented gene-taligene-disease association? Has databas-cocumented gene-disease association?	or protein or Browse] No file selected.
Gene Query Se Cene name or Jripret accession	ACDS category select •	lected AACDS classificatures within a gene Has Righ deviatious count? Has CRU-d-countered gene-haligene-disease association? Has disbase-countered gene-disease association? Has disbase-countered variant-disease association?	or protein] No file selected.
Gene Query Se Gene name or Jriprot accession	ACDS select •	Includ ACOS classificatures within a gene Has high enlansings count? Has DIVIG-documented gene-datagene-datage associato? Has doctases-documented gene-datage associato? Has doctases-documented second-datage associato?	or protein] No file selected.
Gene Query St Gene name or Jripret accession	ACDS category select •	Has high delativities count? Has high delativities count? Has bond-documented gree-datapare disease association? Has documented gree-datapare diseascolation? Has documented gree-datapare diseascolation? Has documented gree-datapare diseascolation?	or protein] No file selected.
Gene Query S Gene name or Jrippet accession AACDS-based Genome Analysis F	AACDS select AACDS select Category select Returns AACDS and rec	Includ AACOS classificatures within a game Nas Rh, Madhithias could take the approximate and the approximate takes controls developed and takes association? Includ controls developed and takes association? Liction score statistics for each AACOS cate	or protein or Browse] No file solected.
Gene Query St Sees name or http://docsaion	ACCS select ACCS selects ACCS selec	Incide AACDS classificatures within a gene Has high advanced gene hatgene-disease substantiation in the action of the advanced set of the substantiation of the advanced set of the advanced set of the substantiation of the advanced set of the advanced set of the substantiation of the advanced set of the advanced set of the substantiation of the advanced set of the advanced set of the substantiation of the advanced set of the advanced set of the substantiation of the advanced set of the advanced set of the substantiation of the advanced set of the advanced set of the substantiation of the advanced set of the advanced set of the substantiation of the advanced set of the advanced set of the advanced set of the substantiation of the advanced set of the adva	or protein or Browse. gory. ut file Browse.] No fire uniterated.
Gene Query Si Cree name or Unyord scossion AACDS-based Genome Analysis Select output format	ACDS select ACDS	lected AACDS classRetures within a gene Has high deletines count? Has both documented gene adoption docume secondar? Has documented gene adoption docume secondar? Has documented gene adoption documented secondar?	or protein or Browse. gory. ut file Browse.] No file selected.
Gene Query St Gene name or ARCD5-based Genome Analysis Select output format	ACCOS category select •	Incred AACDS classificatures within a gene His high distantive count? His Double-manned preh taligene-disease associator? His distance double-taligned disease associator? His distance double-taligned disease His distance double-taligned disease	or protein or Browse. gory. ut fie Browse.) No file welested.

					Н	ome	Report	Statistic	sl
			Associati	on-Ad	justed Consei	nsus D	eleteri	ous Scheme Variant s	(AAC umm
Chromosome	8	Coordinate (hg19)	90993645		Reference nucleotide	G		Alternate nucleotide	A
Uniprot accession	O60934	Amino acid positio	n 93		Reference amino acid	s		Alternate amino acid	L
Gene name	NBN		Gene other	names	NBS; NBS1; P95				
AACDS category	1	Has high delete	tious count?	🖾 Del	eterious predictions	DIPIPIN	DJL	Deleterious co	unt 2
		init generative			e na Gran staat	-			
		Has gene-diseas	e association?	Dise	sase list (gene level)	re br ca ch le	mal cell reakage s incer; ap ildhood sukemia	carcinoma; nijr yndrome; breast lastic anemia; acute lymphoblas (source:	segen stic
		Has variant-dise	ase association?	Dise	ease list (variant level)	ch le aw	ukenia (vissvar)	acute lymphoblas source:	stic
Additional data									
Predictied deleteric	a one	Predicted sequer	ice conservation	scores	Minor allele freque	ncy (ESP65	(00)	Other information	0.04
Debahas 2 HDD/	0.044	00.00 5./10 Date: 0 2.693			Function Attraction	0.05		(present bat	0.65
Polyshan2 HULR	0.461	SiPhy 19.84	4		surveyered American	0.00		(necessary)	0.03
LRT	0.026	10.04							
	0.878								
MutationTaster MutationAssessor	1.545								

- 1							H	ome F	eport	Stat	istics	Hel
				Assoc	iation-	Adjuste	ed Cons	ensus (eleter	ious Scł	heme (/	AACDS
Vhole a												
	enome st	atistics								₩ Do	wnload C	SVID
	enome su	atistics	Average (sl	idev) of delet	arious predi	tion scores		Average (std	ev) of conse	Ve Do	wnload C	IN TX
Category	enome st	SIFT	Average (si Polyphen2 HDIV	odev) of delet Polyphen2 HVAR	arious predia	tion scores Mutation Taster	Mutation Assessor	Average (std GERP	ev) of conse phyloP	Vation score SiPhy	wnload C Average (std AtrAmr	EurAme
Catogory 1, 5	Nariana	SIFT	Average (sl Polyphen2 HDIV 0.77 (0.00)	dev) of delet Polyphen2 HVAR 0.24 (0.00)	LRT	tion scores Mutation Taster 0.00 (0.00)	Mutation Assessor 0.69 (0.00)	Average (std GERP 5.75 (0.00)	ev) of conse phyloP 2.19 (0.00)	Vation score SiPhy 16.07 (0.00)	Average (std AtrAmr 18.54 (0.00)	Evr) of SM EurAm 4.79 (0.0
Category 1, 5 2A	Narians	SIFT 0.57 (0.00) 0.63 (0.29)	Average (si Polyphen2 HDIV 0.77 (0.00) 0.02 (0.03)	Polyphen2 HVAR 0.24 (0.00) 0.04 (0.05)	LRT 1.00 (0.00) 0.93 (0.10)	tion scores Mutation Taster 0.00 (0.00) 0.00 (0.00)	Nutation Assessor 0.69 (0.00) 0.52 (0.04)	Average (std GERP 5.75 (0.00) 3.68 (1.38)	ev) of conse phyloP 2.19 (0.00) 1.65 (0.63)	Vation score SiPhy 16.07 (0.00) 11.23 (4.41)	wnload C Average (std At:Amr 18.54 (0.00) 5.15 (0.47)	EurAme 4.79 (0.0 4.01 (2.2
Category 1, 5 2A 20	Nariants	SIFT 0.97 (0.00) 0.63 (0.20) 0.67 (0.00)	Average (s) Polyphen2 HDIV 0.77 (0.00) 0.02 (0.03) 1.00 (0.00)	Adev) of delet Polyphen2 HVAR 0.24 (0.00) 0.04 (0.06) 0.93 (0.00)	LRT 1.00 (0.00) 0.93 (0.10) 1.00 (0.00)	tion scores Mutation Taster 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	Mutation Assessor 0.59 (0.00) 0.52 (0.04) 0.54 (0.00)	Average (std GERP 5.75 (0.00) 3.68 (1.38) 5.32 (0.00)	ev) of conse phyloP 2.19 (0.00) 1.65 (0.68) 1.60 (0.00)	Do vation score SiPhy 16.07 (0.00) 11.23 (4.41) 14.09 (0.00)	wnload C Average (std AlsAmr 18.54 (0.00) 5.15 (6.47) 1.04 (0.00)	EurAm 4.79 (0.0 4.01 (2.2 6.62 (0.0
Category 1, 5 2A 20 28, 38	Narians 1 3 1 1	SIFT 0.97 (0.00) 0.63 (0.29) 0.67 (0.00) 0.71 (0.00)	Average (al Polyphen2 HDIV 0.77 (0.00) 0.02 (0.03) 1.00 (0.00) 1.00 (0.00)	dev) of delet Polyphen2 HVAR 0.24 (0.00) 0.04 (0.06) 0.93 (0.00) 1.00 (0.00)	LRT 1.00 (0.00) 0.53 (0.10) 1.00 (0.00) 1.00 (0.00)	tion scores Mutation Taster 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	Mutation Assessor 0.59 (0.00) 0.52 (0.04) 0.54 (0.00) 0.52 (0.00)	Average (std GERP 5.75 (0.00) 3.68 (1.38) 5.32 (0.00) 4.49 (0.00)	ev) of conse phyloP 2.19 (0.00) 1.65 (0.60) 1.60 (0.00) 2.49 (0.00)	Do Vation score SiPhy 16.07 (0.00) 11.23 (4.41) 14.69 (0.00) 17.33 (0.00)	wnload C Average (std AtsAmr 18.54 (0.00) 5.15 (6.47) 1.04 (0.00) 0.48 (0.00)	EurAme 4.79 (0.00 4.01 (2.2) 6.62 (0.00 2.35 (0.00
Category 1, 5 2A 20 28, 38 3A	AVariants	SIFT 0.97 (0.00) 0.63 (0.29) 0.67 (0.00) 0.71 (0.00) 0.64 (0.31)	Average (s) Polyphen2 HDIV 0.77 (0.00) 0.02 (0.03) 1.00 (0.00) 1.00 (0.00) 0.40 (0.43)	Adev) of delet Polyphen2 HVAR 0.24 (0.00) 0.04 (0.06) 0.33 (0.00) 1.00 (0.00) 0.31 (0.35)	LRT 1.00 (0.00) 0.53 (0.10) 1.00 (0.00) 1.00 (0.00) 0.54 (0.40)	Mutation Taster 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	Mutation Assessor 0.69 (0.00) 0.52 (0.04) 0.54 (0.00) 0.52 (0.00) 0.59 (0.10)	Average (std GERP 5.75 (0.00) 5.48 (1.38) 5.32 (0.00) 4.49 (0.00) -0.28 (4.75)	ev) of conse phyloP 2.19 (0.00) 1.65 (0.68) 1.60 (0.00) 2.49 (0.00) 0.34 (1.33)	Do Vation score SiPhy 16.07 (0.00) 11.23 (4.41) 14.69 (0.00) 17.33 (0.00) 8.04 (4.27)	wnload C Average (std AtsAmr 18.54 (0.00) 5.15 (6.47) 1.04 (0.00) 0.48 (0.00) 5.48 (8.16)	EurAme 4.79 (0.0 4.01 (2.2 6.62 (0.0 2.35 (0.0 10.80 (15)
Category 1, 5 2A 20 28, 38 3A 5	AVariants AVariants 1 1 1 1 1 1 1 1 1 1 1 1 1	SIFT 0.97 (0.00) 0.63 (0.29) 0.67 (0.00) 0.71 (0.00) 0.64 (0.31) 0.62 (0.33)	Average (s) Polyphen2 HDIV 0.77 (0.00) 0.02 (0.03) 1.00 (0.00) 1.00 (0.00) 0.40 (0.43) 0.49 (0.45)	Adev) of delet Polyphen2 HVAR 0.24 (0.00) 0.04 (0.00) 0.31 (0.35) 0.27 (0.26)	LRT 1.00 (0.00) 0.53 (0.10) 1.00 (0.00) 1.00 (0.00) 0.54 (0.40) 1.00 (0.00)	Mutation Taster 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	Mutation Assessor 0.69 (0.00) 0.52 (0.04) 0.54 (0.00) 0.52 (0.00) 0.52 (0.00) 0.52 (0.00)	Average (std GERP 5.75 (0.00) 3.68 (1.38) 5.32 (0.00) 4.49 (0.00) -0.28 (4.75) 1.23 (1.37)	ev) of conse phyloP 2.19 (0.00) 1.65 (0.68) 1.60 (0.00) 2.49 (0.00) 0.34 (1.33) 0.13 (0.17)	Do vation score SiPhy 16.07 (0.00) 11.23 (4.41) 14.69 (0.00) 17.33 (0.00) 8.04 (4.27) 4.58 (1.37)	wnload C Average (std AtsAmr 18.54 (0.00) 5.15 (6.47) 1.04 (0.00) 0.48 (0.00) 5.48 (8.16) 5.35 (1.46)	EurAme 4.79 (0.00 4.01 (2.2) 6.62 (0.00 2.35 (0.00 10.80 (153 7.84 (1.45



								3	lome	Repo	rt I	Statis	stics	He
				As	sociat	ion-Ad	djuste	d Cor	isensu	s Dele	terio	is Sche	eme (A	ACD
												Vari	iant su	nmar
3ene-by 37 records	gene sta for this sea	tistics rch		То	tal 37 rec	cord(s) <	Previou	s 1 • I	Next >	A		V Dowr	nload C	SV TJ
				Av	erage (stde	v) of delete	rious pre-	liction sco	res	wenge (score	inservation	Average 54	(MDEV) C
Gene name	Uniprot_acc	Category	Warlants	SIFT	Polyphen2 HDIV	Polyphen2 MMR	LRT	Mutation Taster	Mutation Assessor	GERP	phyloP	SiPhy	AtrAmr	EurAn
ADA	P00013	5					1.00 (0.00	1.00 (0.00)	0.46 (0.00)	2.60 (0.00)	0.30 (0.00)	5.96 (0.00)	6.81 (0.00)	
ADA ANKR020A4	Q4UJ75	6	3	0.29 (0.00)	0.00 (0.00) 0.33 (0.47)	0.00 (0.00) 0.47 (0.31)	1.00 (0.00	1.00 (0.00)	0.46 (0.00)	2.60 (0.00) 0.95 (1.05)	0.30 (0.00) 0.22 (0.73)	5.96 (0.00) 5.09 (1.82)	6.81 (0.00)	6.35 (0)
ADA ANKRD20A4 ANKRD20A	Q4UJ75 Q58XX3	5 6 5	3	0.29 (0.00) 0.41 (0.42) 0.95 (0.00)	0.00 (0.00) 0.33 (0.47) 0.98 (0.00)	0.00 (0.00) 0.47 (0.31) 0.00 (0.00)	1.00 (0.00	1.00 (0.00) 0.00 (0.00) 0.00 (0.00)	0.46 (0.00) 0.46 (0.18) 0.58 (0.00)	2.60 (0.00) 0.95 (1.06) -0.15 (0.00)	0.30 (0.00) 0.22 (0.73) -0.04 (0.00)	5.96 (0.00) 5.09 (1.82) 3.21 (0.00)	6.81 (0.00)	6.35 (0) 9.34 (0)
ADA ANKRD20A4 ANKRD30A ANKRD30A	Q4UJ75 Q5BXX3 Q9Y2G4	5 6 5 3A	1 3 1 1 1	0.29 (0.00) 0.41 (0.42) 0.95 (0.00) 0.85 (0.00)	0.00 (0.00) 0.33 (0.47) 0.58 (0.00) 0.72 (0.00)	0.00 (0.00) 0.47 (0.31) 0.00 (0.00) 0.00 (0.00)	0.95 (0.00	1.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	0.46 (0.00) 0.46 (0.18) 0.58 (0.00) 0.60 (0.00)	2.60 (0.00) 0.95 (1.06) 0.15 (0.00) 1.64 (0.00)	0.30 (0.00) 0.22 (0.73) 0.04 (0.00) 0.67 (0.00)	5.96 (0.00) 5.09 (1.82) 3.21 (0.00) 7.71 (0.00)	6.81 (0.00) 3.89 (0.00) 2.42 (0.00)	6.35 (0) 9.34 (0) 13.92 (0
ADA ANKRD20A4 ANKRD30A ANKRD6 ATP6VEA4	Q4UJ75 Q5BXX3 Q59Y2G4 Q5H5G4	5 6 5 3A 1, 5	1 3 1 1 1 1 1	0.29 (0.00) 0.41 (0.42) 0.95 (0.00) 0.85 (0.00) 0.97 (0.00)	0.00 (0.00) 0.33 (0.47) 0.98 (0.00) 0.72 (0.00) 0.77 (0.00)	0.00 (0.00) 0.47 (0.31) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	1.00 (0.00 0.95 (0.00 1.00 (0.00	1.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	0.46 (0.00) 0.46 (0.18) 0.58 (0.00) 0.60 (0.00) 0.69 (0.00)	2.60 (0.00) 0.95 (1.05) 0.15 (0.00) 1.64 (0.00) 5.75 (0.00)	0.30 (0.00) 0.22 (0.73) -0.04 (0.00) 0.67 (0.00) 2.19 (0.00)	5.96 (0.00) 5.09 (1.82) 3.21 (0.00) 7.71 (0.00) 16.07 (0.00)	6.81 (0.00) 3.89 (0.00) 2.42 (0.00) 18.54 (0.00)	6.35 (0) 9.34 (0) 13.92 (0) 4.79 (0)
ADA ANKRD20A4 ANKRD3IA ANKRD6 ATP6V0A4 BTBD11	P00813 Q40375 Q58XX3 Q592G4 Q5918G4 A6QL63	5 6 3A 1, 5 3A	1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	0.23 (0.00) 0.41 (0.42) 0.95 (0.00) 0.85 (0.00) 0.97 (0.00) 0.28 (0.00)	0.00 (0.00) 0.33 (0.47) 0.98 (0.00) 0.72 (0.00) 0.77 (0.00) 0.26 (0.00)	0.00 (0.00) 0.47 (0.31) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	1.00 (0.00) 0.95 (0.00) 1.00 (0.00) 0.68 (0.00)	1.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	0.46 (0.00) 0.46 (0.18) 0.58 (0.00) 0.69 (0.00) 0.69 (0.00) 0.54 (0.00)	2.60 (0.00) 0.96 (1.06) 0.15 (0.00) 1.64 (0.00) 5.75 (0.00) 3.06 (0.00)	0.30 (0.00) 0.22 (0.73) 0.04 (0.00) 0.67 (0.00) 2.19 (0.00) 2.03 (0.00)	5.96 (0.00) 5.09 (1.82) 3.21 (0.00) 7.71 (0.00) 16.07 (0.00) 10.10 (0.00)	6.81 (0.00) 3.89 (0.00) 2.42 (0.00) 18.54 (0.00) 25.82 (0.00)	9.34 (0) 9.34 (0) 13.92 (0) 4.79 (0) 51.14 (0)
ADA ANKRD20A4 ANKRD30A ANKRD30A ANKRD30A ANKRD30A ANKRD30A ANKRD30A ANKRD30A ANKRD30A ANKRD30A ANKRD30A	P00013 Q40375 Q583XX3 Q5972G4 Q5989G4 A6QL63 Q5581V9	5 6 3A 1,5 3A 6	1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	0.23 (0.00) 0.41 (0.42) 0.95 (0.00) 0.85 (0.00) 0.37 (0.00) 0.28 (0.00) 0.28 (0.00)	0.00 (0.00) 0.33 (0.47) 0.58 (0.00) 0.72 (0.00) 0.77 (0.00) 0.26 (0.00) 0.00 (0.00)	0.00 (0.00) 0.47 (0.31) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	1.00 (0.00) 0.95 (0.00) 1.00 (0.00) 0.68 (0.00) 0.50 (0.00)	1.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	0.46 (0.00) 0.46 (0.18) 0.58 (0.00) 0.60 (0.00) 0.69 (0.00) 0.54 (0.00) 0.54 (0.00)	2.60 (0.00) 0.96 (1.06) 0.15 (0.00) 1.64 (0.00) 5.75 (0.00) 3.06 (0.00) 3.62 (0.00)	0.30 (0.00) 0.22 (0.73) 0.64 (0.00) 0.67 (0.00) 2.19 (0.00) 2.03 (0.00) 0.31 (0.00)	5.96 (0.00) 5.09 (1.82) 3.21 (0.00) 7.71 (0.00) 96.07 (0.00) 10.10 (0.00) 2.59 (0.00)	6.81 (0.00) 3.89 (0.00) 2.42 (0.00) 10.54 (0.00) 25.82 (0.00)	9.34 (0) 9.34 (0) 13.92 (0) 4.79 (0) 51.14 (0)
ADA ANKRD30A ANKRD30A ANKRD30A ANKRD30A ANKRD30A ATP\$V0A4 BTBD11 C1orf157 CCL4L1	P00013 Q4UJ75 Q58XX3 Q992G4 Q908G4 A6QL63 Q538V9 Q8NHW4	5 5 3A 1,5 3A 6 3A	1 3 1 1 1 1 1 1 1	0.29 (0.00) 0.41 (0.42) 0.95 (0.00) 0.95 (0.00) 0.97 (0.00) 0.28 (0.00) 0.37 (0.00) 0.31 (0.00)	0.00 (0.00) 0.33 (0.47) 0.58 (0.00) 0.72 (0.00) 0.77 (0.00) 0.26 (0.00) 0.00 (0.00) 0.90 (0.00)	0.00 (0.00) 0.47 (0.31) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	1.00 (0.00 0.95 (0.00 1.00 (0.00 0.68 (0.00 0.50 (0.00 1.00 (0.00	1.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 1.00 (0.00)	0.46 (0.00) 0.46 (0.18) 0.58 (0.00) 0.69 (0.00) 0.54 (0.00) 0.54 (0.00)	2.60 (0.00) 0.96 (1.06) 0.15 (0.00) 1.64 (0.00) 5.75 (0.00) 3.65 (0.00) 3.62 (0.00) 3.62 (0.00)	0.30 (0.00) 0.22 (0.73) 0.67 (0.00) 0.67 (0.00) 2.19 (0.00) 2.03 (0.00) 0.31 (0.00) 1.31 (0.00)	5.96 (0.00) 5.09 (1.82) 3.21 (0.00) 7.71 (0.00) 56.07 (0.00) 50.10 (0.00) 2.59 (0.00) 9.79 (0.00)	6.81 (0.00) 3.89 (0.00) 2.42 (0.00) 18.54 (0.00) 25.82 (0.00) 0.23 (0.00)	6.35 (0) 9.34 (0) 13.92 (0) 4.79 (0) 51.14 (0) 0.50 (0)
ADA ANKRD20A4 ANKRD20A ANKRD20A ANKRD20A ANKRD20A ANKRD20A ANKRD20A ANKRD20A ANKRD20A ANKRD20A ADA ADA ADA ADA ADA ADA ADA ADA ADA A	P00073 Q4UJ75 Q58XX3 Q972G4 Q918G4 AEQL63 Q55NV9 Q8180W4 P25024	5 5 3A 1,5 3A 6 3A 3A	1 3 1 1 1 1 1 1 1 1 2	0.29 (0.00) 0.41 (0.42) 0.95 (0.00) 0.85 (0.00) 0.97 (0.00) 0.97 (0.00) 0.91 (0.00) 0.94 (0.00) 0.96 (0.34)	0.00 (0.00) 0.33 (0.47) 0.58 (0.00) 0.72 (0.00) 0.77 (0.00) 0.26 (0.00) 0.00 (0.00) 0.90 (0.00) 0.49 (0.49)	0.00 (0.00) 0.47 (0.31) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00)	1.00 (0.00 0.95 (0.00 1.00 (0.00 0.68 (0.00 0.50 (0.00 0.03 (0.03	1.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 1.00 (0.00) 0.00 (0.00)	0.46 (0.00) 0.46 (0.18) 0.58 (0.00) 0.69 (0.00) 0.69 (0.00) 0.54 (0.00) 0.43 (0.00) 0.43 (0.00)	2.60 (0.00) 0.96 (1.06) 0.15 (0.00) 1.64 (0.00) 5.75 (0.00) 3.06 (0.00) 3.62 (0.00) 3.62 (0.00) 3.62 (0.00) 3.95 (6.75)	0.30 (0.00) 0.22 (0.73) 0.67 (0.00) 0.67 (0.00) 2.19 (0.00) 2.03 (0.00) 0.31 (0.00) 1.31 (0.00) 1.31 (0.00)	5.96 (0.00) 5.09 (1.82) 3.21 (0.00) 7.71 (0.00) 56.07 (0.00) 10.50 (0.00) 2.59 (0.00) 5.79 (0.00) 4.57 (4.41)	6.81 (0.00) 3.89 (0.00) 2.42 (0.00) 18.54 (0.00) 25.82 (0.00) 0.23 (0.00) 3.04 (2.47)	6.35 (0) 9.34 (0) 13.92 (0) 4.79 (0) 51.14 (0) 0.58 (0) 3.20 (0)
ADA ANKRD20A4 ANKRD20A4 ANKRD30A ANKRD30A ANKRD30A BTBD11 C1ort197 CCL4L1 CXCR1 CXCR1	P00073 Q4UJ75 Q38XX3 Q992G4 Q9919G4 A6QL63 Q558V9 Q091904 P25024 P11509	5 5 3A 1, 5 3A 6 3A 3A 3A 3A	1 3 1 1 1 1 1 1 1 2 1 1	0.29 (0.00) 0.41 (0.42) 0.95 (0.00) 0.85 (0.00) 0.97 (0.00) 0.28 (0.00) 0.97 (0.00) 0.91 (0.00) 0.56 (0.34) 1.00 (0.00)	0.00 (0.00) 0.33 (0.47) 0.58 (0.00) 0.72 (0.00) 0.77 (0.00) 0.26 (0.00) 0.00 (0.00) 0.49 (0.49) 0.09 (0.00)	0.00 [0.00] 0.47 [0.31] 0.00 [0.00] 0.00 [0.00] 0.00 [0.00] 0.00 [0.00] 0.00 [0.00] 0.49 [0.12] 0.00 [0.00]	1.00 (0.00 0.95 (0.00 1.00 (0.00 0.68 (0.00 0.50 (0.00 1.00 (0.00 0.03 (0.01 0.99 (0.01	1.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 0.00 (0.00) 1.00 (0.00) 1.00 (0.00) 0.00 (0.00)	0.46 (0.00) 0.46 (0.18) 0.58 (0.00) 0.69 (0.00) 0.59 (0.00) 0.54 (0.00) 0.43 (0.00) 0.52 (0.06) 0.76 (0.00)	2.60 (0.00) 0.96 (1.06) 1.64 (0.00) 5.75 (0.00) 3.62 (0.00) 3.62 (0.00) 3.62 (0.00) 3.62 (0.00) 3.63 (6.75) 2.77 (0.00)	0.30 (0.00) 0.22 (0.73) 0.04 (0.00) 0.67 (0.00) 2.19 (0.00) 2.03 (0.00) 0.31 (0.00) 1.31 (0.00) 1.07 (1.69) 1.14 (0.00)	5.96 (0.00) 5.09 (1.82) 3.21 (0.00) 7.71 (0.00) 16.07 (0.00) 2.59 (0.00) 9.79 (0.00) 4.57 (4.41) 9.86 (0.00)	6.81 (0.00) 3.89 (0.00) 2.42 (0.00) 18.54 (0.00) 25.82 (0.00) 0.23 (0.00) 3.04 (2.47) 0.45 (0.00)	6.35 (0) 9.34 (0) 13.92 (0) 4.79 (0) 51.14 (0) 0.50 (0) 3.20 (0) 2.47 (0)

		aren			Total 5 ro	cord(e) <	Provinue 1	- Next >	P Dow	midad Cov 1X
Position	RSid	Gene name	AA SAMAF	EA SAMAE	Category	Deleterious count	Deleterious	Trait list	Disease list	Disease list
8:90990521 T.:>A ((171 F)	1361754966	MBN			1, 29	5	DEDEDEDEM	-	renal cell carcinoma; nijmegen breekoge syndrome; breast cascer; aplastic anemia; childhood acute lymphoblastic leukemia	aplastic anemia (source msv3d)
8:90990521 T>C () 171 V)	1561754866	MSN	0.05	0.16	1,28	4	DIDIDINAL	-	renal cell carcinoma; njimegen breakope syndrome; breast cancer, ajdastic anemia; chidhood acate lymphoblastic leskenia (source: swisover)	childhood acute Iymphoblastic loukemi (source: swissear)
8:90990521 Т.эб () 171 L)	rs61754866	MBN			1, 28	5	OLOUDIDIM	-	renal cell carcinoma; njimegen breakope syndrome; breast cancer, ajlastic anemia; childhood acate lymphoblastic leskemia (source: swisover)	aplastic anemia (sourc msv3d)
8:90992994 G.⇒A (L 150 F)		MIM			1, 28	5	DIDIDIDIDI	-	renal cell carcinoma; njimejen breakoje syndrome; breast cancer, algastic anemia; chidhood acate lymphoblastic leakemia (soerce: swissver)	breast cancer (bc) (source: msv34)
8:90993540 C.→T (D 95 N)	rs61753720	NBN	0.55	0.25	1, 28	4	DEDIDINIDIM	-	renal cell carcinoma; nijmogen breakogo syndrome; breast cancer; aplastic anemia; childhood acute lymphoblastic leukemia	childhood acute hymphoblastic leukemi (source: swisswar)

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 <u>http://cig.gatech.</u> website: (1) Variant query, (2) Gene query, and <u>edu/tools</u>. All standard browsers are supported. (3) AACDS-based genome analysis. Users can

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.

e780

Research Papers

Queries	Query options	File	formate	6		Column descriptions
	By DNA	Chr:10 2678125 Chr:10 2678125 Chr:10 2678125	7 7 7	T T T	A C G	1 = chromosome number 2 = hg19 coordinate 3 = reference nucleotide 4 = alternative nucleotide
Variant query	By protein (providing gene or protein names)	Gene:AACS Gene:GOT1 Gene:NT5C2 Or Uniprot:Q8IZY2 Uniprot:Q86UK0	8 413 515 2000 2000	G Q K N T	S H Q K A	 1 = gene name or UniProtKB accession number 2 = amino acid position (UniProtKB numbering) 3 = reference amino acid 4 = alternative amino acid
		Uniprot:095477	2000	L	R	
Gene query	Providing gene or protein names	Gene:HSD3B2 Gene:ABCA12 Gene:SH3BP2 Or	1 1 1			1 = gene name, or UniProtKB acces- sion number 2 = AACDS category (1, 2A, 2B, 3A, 3B, 4, 5, or 6)
		Uniprot:Q9NY61	4 2B 4			
AACDS-based genome analysis	-	Chr:10 2678125 Chr:10 2678125 Chr:10 2678125	57 57 57	T T T	A C G	1 = chromosome number 2 = hg19 coordinate 3 = reference nucleotide 4 = alternative nucleotide

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.

the AACDS classification of their variant of interest by providing some search parameters: for query by DNA, chromosome number, ha19 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

For a single-entry query, users can search for 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

EMBnet.journal 20

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. <u>http://dx.doi.org/10.1002/ humu.21260</u>
- 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. <u>http://dx.doi.</u> 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 diseasegene 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. <u>http://dx.doi.org/10.1038/</u> <u>nmeth.3046</u>
- 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. <u>http://dx.doi.org/10.1038/</u> 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. <u>http://</u> 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. <u>http://</u> 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. <u>http://dx.doi.</u> 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. <u>http://dx.doi.org/10.1371/journal.</u> 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. <u>http://</u> 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. <u>http://dx.doi.org/10.1186/1756-0381-6-24</u>
- 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. <u>http://dx.doi.org/10.1038/nmeth.2656</u>
- 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. <u>http://dx.doi.org/10.1016/j.ajhg.2012.10.015</u>

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.