# Supplementary material for Part XY (ABC lab analysis)

This page summarizes protective variants observed in the Zimmerome as determined by Ali Torkamani’s group at the Scripps Translational Science Institute

**Continue reading 🡪**

Schema of data generation and analyses are summarized in the following presentation: **torkamani\_overall\_schema.pptx**

## 1) Variant Annotation

**A slide providing a visual overview of variant annotation:** [**torkamani\_annotation\_schedma.ppt**](http://archive.gersteinlab.org/proj/zimmerome/part05/ppt/Slides.Variant_Calling-Examples_FN.pdf)**x**

### a) Variant Calling

Variant calls were provided to the Torkamani group. Typically the Torkamani group performs variant calling via Genome Analysis Toolkit Best Practices (<http://www.broadinstitute.org/gatk/guide/best-practices.php>). Variant quality scores are passed to the annotation step in order to retain only high quality variants.

### b) Variant Annotation

Variant annotation was performed using the SG-ADVISER system. A detailed description is provided below:

TOOL: [**genomics.scripps.edu/ADVISER**](https://genomics.scripps.edu/ADVISER/)

SG-ADVISER requires, as input, only a list of variants including chromosome, start and end position, and alleles. Variant submission sets off a cascade of computational processes, dispatched by an automated task scheduler, that drives variants through the basic annotation workflow presented. SG-ADVISER is capable of analyzing single nucleotide variants, insertions, deletions, and block substitutions.

SG-ADVISER produces four major types of annotations: 1) annotation of the genomic element within which a variant resides (e.g., exon, promoter, conserved element, transcription factor binding site, protein domain etc.); 2) prediction of the functional impact of a variant on a genomic element (conservation level, prediction of impact on protein function, changes in transcription factor binding strength, splicing efficiency, microRNA binding, etc.); 3) annotation of molecular and biological processes which link variants across genes and/or genomic elements with one another, and 4) annotation of known clinical characteristics of the gene or variant (e.g. pharmacogenetic variants, GWAS associations, eQTLs etc.) and prediction of clinical characteristics due to novel variants predicted to damage known disease associated genes. Specific annotations relevant to this analysis are as follows:

1) Variants are mapped to known genes and their impact on gene function is called (nonsynonymous, synonymous, frameshift, nonsense, splice site donors, splice site acceptors, microRNA binding sites, transcription factor binding sites etc.). Gene models are based upon the UCSC known genes track for hg19 (Fujita et al. 2011). Coding changes are calculated based upon a custom script. Splice site donor and acceptor sites are the first and last two bases of an intron. Transcription factor binding sites are based upon scanning of TRANSFAC transcription factor motifs (Wingender et al. 1996) against the entire genome, and filtered stringently in non-DNAse hypersensitive, non-conserved element, or non-promoter regions (Siepel et al. 2005, Raney et al 2011). microRNA binding sites are determined based upon TargetScan mapping of known microRNAs to the 3’utrs of known genes (Lewis et al. 2005).

2) Variants are mapped to dbSNP ID’s or other non-ID assigned but previously observed variantions, and their observed frequency in reference populations are determined. Population frequencies are based upon the 1,000 genomes data, 69 publically available Complete Genomics Genomes, and ~450 Wellderly Genomes, genomes of individuals over the age of 80 with no common chronic conditions, also sequenced by Complete Genomics are extracted.

3) Variants are subject to functional prediction. Nonsynonymous variants are analyzed for their impact on coding genes using the Polyphen-2 (Adzhubei IA et al. 2010), SIFT (Sim et al. 2012) and Condel (González-Pérez and López-Bigas 2011) algorithms for nonsynonymous variants. Prediction of damaged truncated genes by the proportion of conserved bases removed by the truncation, where approximately >4% of the conserved portion of a gene must be removed by a truncation as determined by determining C-terminal truncation or N-terminal loss before the next available start codon (Hu and Ng, 2012). In-frame substitutions are analyzed using the Log.R.E-value approach with HMMR (Clifford et al. 2004). Mutations impacting splice site donors or acceptors are identified. Variants nearby splice sites, but not impacting the first and last two bases of an intron are analyzed by the MaxENT algorithm (Yeo and Burge 2004). Transcription factor binding site perturbation is called based upon changes in motif matching score at the site of the variant (Andersen et al 2008). Lost microRNA sites are determined via TargetScan of the variant site (Lewis et al. 2005).

4) Known disease causing mutations and disease associated genes are extracted from Online Mendelian Inheritinance in Man (McKusick et al. 1998) and the Human Gene Mutation Database (Stenson et al. 2012). False positives are removed by filtering on allele frequency.

5) A slightly modified American College of Medical Genetics (ACMG) classification is performed by combining all of the above annotations (Richards et al. 2008). For the most part, SG-ADVISER annotations conform to ACMG categories with a few exceptions: 1) reported disease causative variants are only considered true category 1 variants if their allele frequency is <1%. This filter removes the large number of false positives known to inhabit disease mutation databases. Reported disease causative variants of 1-5% allele frequency are placed in category 2 – which is formally a category reserved for variants unreported but expected to be causative for a disease. Category 2 is filtered for truncation mutations that are not predicted to be damaging, that is, frameshift or nonsense variants towards the beginning or end of a gene. The cutoff giving the greatest accuracy is a requirement that >4% of the conserved portion of the affected gene must be removed by the truncation. However, in order to capture truncating mutations that do not meet this threshold, but are still potentially interesting, we created a 2\* category to maintain and highly rank these sorts of variants. A full description of the requirements for each category is available on the SG-ADVISER website.

### c) Variant Filtration

Variant filtration was performed using a custom script. The output can be found in the file below.

***associated data files: .***

**PG0004515\_raw\_filtered\_annotation.xlsx**

Variants were filtered to retain high quality SG-ADVISER class 1, 2, and 6 variants (i.e. variants known or predicted to be functional, and those associated with various conditions – as described above). The approximately 1,500 variants retained can be found in the above file on the sheet “Raw.”

These variants were then filtered further in order to only retain variants where the evidence description contained keywords such as “protective”, “resistance” etc. The 75 variants retained during this step can be found in the above file on the sheet “Filtered”. The literature evidence supporting each variant was then reviewed and the final report generated based on those deemed to be high confidence fidings by expert review. Variant filtration can also be performed in the SG-ADIVSER User Interface. (http://genomics.scripps.edu/ADVISER/downloads.jsp).

***References :***

Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248-9.

Andersen MC, et al. In silico detection of sequence variations modifying transcriptional regulation. PLoS Comput Biol. 2008 Jan;4(1):e5.

Clifford RJ, Edmonson MN, Nguyen C, Buetow KH. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics. 2004 May 1;20(7):1006-14.

Fujita PA, et al. The UCSC Genome Browser database: update 2011. Nucleic Acids Res. 2011 Jan;39(Database issue):D876-82.

González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011 Apr 8;88(4):440-9.

Hu J, Ng PC. Predicting the effects of frameshifting indels. Genome Biol. 2012 Feb 9;13(2):R9.

Jeh TH et al. An Analytical Comparison of Approaches to Personalizing PageRank. 2003

Lewis BP, et al. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005 Jan 14;120(1):15-20.

Lohmueller KE et al. Proportionally more deleterious genetic variation in European than in African populations. Nature 2008 451:994-7.

McKusick VA. Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. Baltimore: Johns Hopkins University Press, 1998 (12th edition).

Raney BJ, et al. ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res. 2011 Jan;39(Database issue):D871-5.

Richards CS, et al. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med. 2008 Apr;10(4):294-300.

Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug;15(8):1034-50.

Sim NL, et al. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W452-7.

Stenson PD, et al. The Human Gene Mutation Database (HGMD) and Its Exploitation in the Fields of Personalized Genomics and Molecular Evolution. Curr Protoc Bioinformatics. 2012 Sep;Chapter 1:Unit1.13.

Wingender E, et al. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. 1996 Jan 1;24(1):238-41.

Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004;11(2-3):377-94.

Zhang H, et a. Phenotype-information-phenotype cycle for deconvolution of combinatorial antibody libraries selected against complex systems. Proc Natl Acad Sci U S A. 2011 Aug 16;108(33):13456-61.