

## Supplemental Data S1.

### 1 WHOLE PROTEIN MODELING USING I-TASSER AND VISUALIZATION WITH JSmol

We employed the protein structure prediction software suite, I-TASSER (Roy, et al., 2010; Yang, et al., 2013; Zhang, 2008) Vers. 3.0, 2.1 and 1.1) to create full-length models from amino acid sequences. Protein transcript input sequences for each gene were downloaded in FASTA format from the NCBI GenPept database ([www.ncbi.nlm.nih.gov/protein](http://www.ncbi.nlm.nih.gov/protein)). The protein modeling calculations were carried out using a locally installed copy of I-TASSER with a few minor modifications for server compatibility. A small number of models were also built using the I-TASSER online server (<http://zhanglab.ccmb.med.umich.edu/I-TASSER>). I-TASSER models were built using the default software options with the following exceptions: the sequence identity cut-off was set to 0.3, the number of top template outputs for each threading program was set to 15, the number of final models was set to 3, and the parallel mode was invoked for the computation. For I-TASSER's BLAST sequence comparisons, a local copy of the BLAST database was used. Protein transcripts larger than 1500 amino acid residues could not be modeled due to I-TASSER software's limitation. The completed models for each protein transcripts were manually inspected using JSmol (<http://jsmol.sourceforge.net/>), a HTML5 and non-Java implementation of Jmol. JSmol was used on the AVIA website for visualization. Custom JSmol scripts were developed for optimal webpage visualization.

For a complete list of modeled protein structures, please visit <http://avia.abcc.ncifcrf.gov/apps/site/struct>. On the AVIA website, the users will be presented with a fully functional JSmol viewer. The default right-click option can be invoked to change the protein display or export it to a file and the advanced users can take advantage of the command console to create custom visualizations. The coloring scheme used for secondary structure display can be found on the Jmol website, <http://jmol.sourceforge.net/jscolors/#Secondary%20structure>. For each non-synonymous SNP in the user annotated list, a custom script retrieves the available model(s) from the local AVIA structure library and displays them to the user. The amino acid position affected by the mutation is indicated in the model as red colored atomspheres. The chemical structures of the original and variant amino acid residues are also shown in a side-by-side comparison at the bottom of the results webpage.

If an in-house full length structure is yet modeled, the PDB id and mapped protein position is obtained using the Uniprot API. If the protein position falls within the publicly available structure, it is downloaded from PDB and used for visualization. This feature is available through the "Feature annotation and Visualization", "Cascade Filtering", and "Visualization of Protein using JSmol" workflows.

### 2 WEB DISPLAY LIMITATIONS

Some users have experienced issues due to server load time for large datasets. To address the lag issue for larger datasets on the web, some results are pre-filtered and displayed on the results page so that users can still take advantage of the gene-based tools on the website. If the original variant file contains more than 5,000 mutations, the AVIA results page displays the pre-filtered results as follows:

1. On the "Variant Annotations" tab, the variant annotation table will only show variants with variants with AVIA summary code categories (described in Section 5 and Table 2).
2. On the "Protein Features" tab, the protein variant results from Uniprot will show the first 100 protein annotations.
3. On the "David gene clustering" tab, the gene list uploaded to the DAVID server will only have genes that have a variant that have been designated with a code in the "AVIA Summary" column of the Variant Results. If there are more than 2000 genes in selected gene list, then it will disable submissions to the DAVID server due to their limitations.
4. On the "Visualization" tab and jsmol visualization page, protein variants with available models have been preselected for the user.

All data will be available for download to the user. Pages which do not display the users' complete dataset are identified in the summary section of each tab.

### 3 IMPLEMENTATION OF FUNSEQ2

FunSeq2 (Fu, et al., 2014) is a variant prioritization algorithm used to score both coding and non-coding mutations for cancer variant sets. FunSeq v2.1 executable was downloaded from <http://funseq2.gersteinlab.org/downloads>, as well as the associated data. Outputs for the SNP and indel pipeline were merged by headers into one field where available; therefore, if

there were no indels in your input variants, there will be no headers for indels. For AVIA, the default parameters were used for analysis with the personal genome option (-m2). FunSeq2 is available upon users' request on the submission page through the "Feature annotation and Visualization" and "Cascade Filtering" workflows under the "Prioritization" section.

#### 4 VARIANT ANNOTATION SUMMARY COLUMN

AVIA automatically generates a summary column for each variant in the user's input based on the databases selected for annotation. The summary column will help users prioritize variants of interest. Table 2 describes the how the summary codes are assigned to each variant.

Table 1. Assignment of Summary Codes.

Code	Assignment of One Letter Codes
<b>D</b>	If a variant was annotated as damaging in two or more of the protein scoring algorithms, SIFT, Polyphen2, Mutation Taster, and Mutation Assessor. For Mutation Taster and Mutation Assessor, they may be annotated as "disease causing" and "Med"/"High", respectively.
<b>P</b>	If a variant was in a site where a post-translational modification occurs
<b>C</b>	If a variant was found in the COSMIC database
<b>F</b>	If a variant had a score of 2 or higher as determined by FunSeq2 prioritization score (coding or non-coding)
<b>O</b>	If a variant is in a gene that is flagged by OMIM
<b>V</b>	If a variant had a hit in the ClinVar database and was annotated as "pathogenic"

As stated in section 2, due to server constraints, variant datasets with more than 5000 variants will only display those variants with a non-empty value in the "Summary" column. However, users will be able to download full annotations in their final reports. The variant summary column is available through the "Feature Annotation and Visualization" and "Cascade Filtering" workflows.

#### 5 GENE SUMMARY BY CATEGORIES

AVIA automatically generates a summary gene list based the user's gene list and the genic variants' specific characteristics, e.g., variants that cause a damaging call in the protein scoring algorithms or that are clinically significant, which are designated as affected using a simple "Y" or "N" state. The summary assessments are dependent on the databases that the user selects upon submission. i.e., if a user does not select FunSeq2, its impact will not be evaluated. All coding and non-coding genes are used for assessment; however, only genes with at least one category affected are displayed. In addition to the category summary, counts for the number of non-synonymous mutations per gene (#NS) and counts of the number of categories with hits (#DB) are included for sorting. If Ensembl annotations were selected at the time of submission, an additional column with the Gene symbols will be displayed. Table 3 describes the categories and their databases, as well as how an affected state is designated; each download archive will include a similar table, as these designations may change over time with the addition of more databases or categories. This feature is only available through the "Feature Annotation and Visualization" and "Cascade Filtering" workflows.

Table 2. Categories, Databases, and Designation of Affected State for Prioritized Gene Lists

Category (Abbreviation)	Databases	Designation of Affected State
<b>Clinical (CL)</b>	ClinVar (Landrum, et al., 2014)	If a gene has a variant with annotation that is labeled "pathogenic" in the database
<b>Disease associated (DI)</b>	COSMIC(Forbes, et al., 2008), OMIM ( <a href="http://www.ncbi.nlm.nih.gov/omim">http://www.ncbi.nlm.nih.gov/omim</a> ), KEGG Disease ( <a href="http://www.genome.jp/kegg/disease/">http://www.genome.jp/kegg/disease/</a> )	If a gene or variant within the gene has a hit in any database in category
<b>Encode (ENC)</b>	Any ENCODE database	If a gene has a variant with hit in any database
<b>FunSeq2 Prioritization (FS)</b>	FunSeq2	If a gene has a variant with a score >2 in coding or non-coding score
<b>Genomic Variants (DGV)</b>	Database for Genomic Variants(MacDonald, et al., 2014)	If a gene has a variant with a hit in any database in category

<b>Interactions (PPI)</b>	FunSeq2_network.hub	If a gene has a variant with a hit in any database in category
<b>KEGG Disease (KD)</b>	KEGG Disease	If a gene has a variant with a hit in any database in category
<b>Motifs (MOT)</b>	ESEFinder(Cartegni, et al., 2003), transcription factor binding site, VISTA	If a gene has a variant with a hit in any database in category
<b>Multiple Variants (MDV)</b>	User Data	If a gene has multiple variants with 2+ damaging calls from any protein scoring algorithms
<b>ncRNA (NC)</b>	Lncipedia (Volders, et al., 2013), miRBase (Griffiths-Jones, et al., 2008)	If a gene has a variant with a hit in any database in category
<b>Mendelian Inheritance in Man (MIM)</b>	Online Mendelian Inheritance in Man	If the gene is in OMIM
<b>Protein Scoring Algorithms (PSA)</b>	SIFT (Kumar, et al., 2009; Sim, et al., 2012), Polyphen2 (Adzhubei, et al., 2013), Mutation Taster (Schwarz, et al., 2010), Mutation Assessor (Gnad, et al., 2013), Provean (Choi, et al., 2012)	If a gene has at least one variant with >=2 annotations that is "damaging", "disease causing", "deleterious", or "medium/high" prediction
<b>Post Translational Modifications (PTM)</b>	PhosphoSite (Hornbeck, et al., 2012), Phosida (Gnad, et al., 2011)	If a gene has a variant with a hit in any database in category
<b>Splicing (SPL)</b>	Ensembl alt splice, TASSDB(Hiller, et al., 2007)	If a gene has a variant with a hit in any database in category
<b>Targets of ncRNA (TNC)</b>	HMDD (Lu, et al., 2008), microPIR_targets (Piriyapongsa, et al., 2012), targetScan (Grimson, et al., 2007; Lewis, et al., 2005), wgRNA	If a gene has a variant with a hit in any database in category
<b>Zygoty (ZYG)</b>	User Data (from VCF file only)	Zygoty is extracted from user's VCF file and any homozygous variant is considered a hit

## 6 PATHVIEW IMPLEMENTATION

PathView (Luo and Brouwer, 2013) v1.2.3 is an R package (v3.0.1) that renders biological data on top of KEGG pathway maps. Using custom perl scripts, AVIA annotations are passed into PathView with transformed scores ranging from 0 to 1, where 0 represents no or unknown impact and 1 represents high impact. Only genes with mutations affecting coding regions or splice sites were used in this analysis. Annotations with no hits or with hits of no significance in their respective databases are assigned a value of 0. Databases that were not selected for annotation have a value of 'NA'. In AVIA v2.0, the default databases used as the pathway states are SIFT, Polyphen-2 (using the human variations set), Mutation Taster, Mutation Assessor, FunSeq2 and COSMIC. These databases were selected because they each ascribe importance to genes (based on the mutation or gene function); future implementations may allow users to choose their own databases. Except for COSMIC, the default databases have a score which can easily be used in PathView. Table 4 describes how the values are transformed for each protein scoring and disease-causing for PathView visualization.

**Table 3. Database and Description of Transformed values for PathView Integration**

Database	Description of Transformed PathView annotation Scores
<b>SIFT</b>	SIFT scores are between 0 and 1, all mutations with scores were used. For SIFT v68, scores were subtracted from 1 since the most damaging scores in SIFT are closer to 0. For ljb26_sift, the scores are left unchanged as this normalization has already been performed.
<b>Polyphen2, Mutation Taster</b>	Scores are not transformed as the scores are between 0 and 1
<b>Mutation Assessor</b>	Scores for Mutation Taster ranged from negative values to positive values. Predictions were transformed into a value used in PathView: Neutral=0.1 Low=0.3 Medium=0.6 High or Deleterious=1.0
<b>COSMIC</b>	Binary score for presence (1) or absence (0) of the specific mutation in the database for a mutation in that gene
<b>FunSeq2 Prioritization score</b>	FunSeq2 coding or noncoding score is divided by 5 to get a number between 0 and 1.

If multiple mutations exist within the same gene with varying scores, the most severe value from each database is used. Using bioDBnet's db2db, associated KEGG pathways in which the genes are involved were found. Some pathways involved could not be rendered using PathView due to the data type and are not displayed to the user. The PathView feature is available upon user's request through the "Feature Annotation and Visualization" and "Cascade Filtering" workflows.

## 7 USE OF REFERENCE TISSUE EXPRESSION DATA FROM GNF

BioGPS (Wu, et al., 2009), a gene annotation portal, contains data ranging from gene ontology to functional annotation. It also contains reference expression gene levels across various tissue types available through the Gene Atlas. For each gene, these expression levels (download date Mar 2013) were ranked by tissue type and then divided into three groups; high, medium, and low expression based on their values. If a gene had multiple probes, the highest expression level amongst the probes was used for each tissue type. The relative expression level (high, medium, or low) for each tissue type is displayed on the interactive results page for the genes present in the user's variation dataset. If the data was run through the "Feature Annotation and Visualization" workflow, a count of the number of mutations in that gene with damaging predictions, assuming that the users selected predictive scoring algorithms for annotation, is also shown; otherwise, it is left blank. Users can further explore the original data by clicking on the gene name and the BioGPS normal expression graph for each tissue type is displayed. By default, this feature is available through the annotation workflows, "Feature Annotation and Visualization", "Cascade Filtering", "Gene based annotation" and "Annotation with Protein coordinates".

## 8 IMPLEMENTATION DIFFERENCES BETWEEN AVIA V1.0 AND AVIA V2.0

In the first version of AVIA, we leveraged ANNOVAR as the foundation for gene based annotations. AVIA v2.0 builds upon its original implementation by incorporating a Flexible Database Integration (FDI) layer, which are utility functions that help to integrate information stored in relational databases. (For more information about FDI, please refer to <http://biodbnet.abcc.ncifcrf.gov/dbInfo/fdi.php>.) On the back end of the application, there is a configurable xml file that can execute queries in Oracle and invoke scripts, thus minimizing maintenance on the website when additional features are added. Also, bioDBnet, a conversion tool that helps associate disparate identifiers, was implemented so users can readily search disparate databases.

From the users' perspective, the look and feel of the results page has changed to highlight various aspects of gene regulation and cell function. Previously, to reduce server load time, we only displayed annotation of genes with damaging predictions in both SIFT (Kumar, et al., 2009; Sim, et al., 2012) and Polyphen2 (Adzhubei, et al., 2013); however, we have improved this by adding javascript libraries for fully functional search capabilities within their original report. As before, users can download all data to their desktops.

## REFERENCES

- Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* 2013; Chapter 7: Unit 7.20.
- Cartegni, L., et al. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res* 2003; 31(13): 3568-3571.
- Choi, Y., et al. Predicting the functional effect of amino acid substitutions and indels. *PLoS one* 2012; 7(10): e46688.
- Forbes, S.A., et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* 2008; Chapter 10: Unit 10.11.
- Fu, Y., et al. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome biology* 2014; 15(10): 480.
- Gnad, F., et al. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC genomics* 2013; 14 Suppl 3: S7.
- Gnad, F., Gunawardena, J. and Mann, M. PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res* 2011; 39(Database issue): D253-260.
- Griffiths-Jones, S., et al. miRBase: tools for microRNA genomics. *Nucleic Acids Res* 2008; 36(Database issue): D154-158.
- Grimson, A., et al. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell* 2007; 27(1): 91-105.

Hiller, M., *et al.* TassDB: a database of alternative tandem splice sites. *Nucleic Acids Res* 2007;35(Database issue):D188-192.

Hornbeck, P.V., *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 2012;40(Database issue):D261-270.

Kumar, P., Henikoff, S. and Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* 2009;4(7):1073-1081.

Landrum, M.J., *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014;42(Database issue):D980-985.

Lewis, B.P., Burge, C.B. and Bartel, D.P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005;120(1):15-20.

Lu, M., *et al.* An analysis of human microRNA and disease associations. *PloS one* 2008;3(10):e3420.

Luo, W. and Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics (Oxford, England)* 2013;29(14):1830-1831.

MacDonald, J.R., *et al.* The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 2014;42(Database issue):D986-992.

Piriyaopongsa, J., *et al.* microPIR: an integrated database of microRNA target sites within human promoter sequences. *PloS one* 2012;7(3):e33888.

Roy, A., Kucukural, A. and Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols* 2010;5(4):725-738.

Schwarz, J.M., *et al.* MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods* 2010;7(8):575-576.

Sim, N.L., *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* 2012;40(Web Server issue):W452-457.

Volders, P.J., *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res* 2013;41(Database issue):D246-251.

Wu, C., *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* 2009;10(11):R130.

Yang, J., Roy, A. and Zhang, Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics (Oxford, England)* 2013;29(20):2588-2595.

Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics* 2008;9:40.