**GESPA Dataset Testing Instructions**

SUNY Upstate Medical University

The GESPA package includes preconfigured batch files for the humsavar dataset. The following instructions describe how to run the dataset, which displays GESPA’s capabilities. To run the humsavar dataset, it is recommended that a system with at least 4 GB or RAM and a dual core processor is used with 64 bit Java installed.

1. Ensure that GESPA is installed property (see user guide).
2. Open GESPA and verify that the program is connected to the internet, the cloud, and both Java and GESPA are up to date. See troubleshooting in the user guide if issues arise. Select the use PSIC score to predict pathogenicity checkbox.

**Note:** Using the PSIC Score to predict the pathogenicity can increase computational times and is not necessary to test data. The data in the Sample\_Data folder was run with the PSIC score enabled.

1. Press the New Batch button and choose Batch1 from the file chooser in the Humsavar\_Test\_Batches folder in the GESPA package. The batch should begin automatically.
2. Create 2 text files entitled align2 and align3 in the same folder. Also in this folder, make a copy of the align1\_Sample from the Sample\_Data folder in the GESPA package.
3. Once the batch has completed, from the Batch Summary page, press Cons. Report to open the conservation report.
4. In the Batch summary page press alt-a followed by control-c (command c on mac) to select and copy all of the data. Paste this data into the align2 text file. Follow the same process for the conservation report except using the align3 file. Save both files.

**Note:** It is recommended that the files are saved in the ANSI format.

1. Close the program and repeat steps 2-6 with Batch2-Batch11.
2. Open the GESPA\_Data\_Alignment\_Tool folder. Open the AlignData program.

**Note:** The source code for the AlignData Jar file is available in the GESPA\_Data\_Alignment\_Tool folder. The program can be installed from the source in a similar method to GESPA except no external JAR files are needed.

1. In the GESPA Data Alignment Tool select the buttons for Batch Text, Gene Info Panel. And Cons. Report. When the file chooser opens, choose the align1\_Sample, align2, and align3 files respectively.

**Note:** The GESPA Data Alignment Tool can be used with regular batch files in a similar method. Instead of the align1\_Sample file, a batch file with annotations can be used (provided that the Gene and the SNP are the first two columns) and the align2 and align3 files can contain the data from the program information panel and conservation report generated from running the batch file.

1. Press the Align Datasets button and copy the output from the program’s text box into a spreadsheet program (ex. Excel). The columns can be split and reformatted in spreadsheet program. See the Full\_Data.xlsx file in the Sample\_Data folder in the GESPA package for an example.

**Note:** The Full\_Data.xlsx file contains 3 sheets. The Full Data, Adjustable sheet contains all of the data from the above method as well as adjustable formulas which display the accuracy of GESPA predictions for user-entered PSIC and WPC cutpoints. The Condensed Data, Adjustable sheet contains a condensed but still user-adjustable version of the Full Data sheet. Finally the Adjustable, no Lit sheet contains the same information as the Full Data sheet but does not take into account the literature based predictions of GESPA.

1. Once the data above is gathered further analyses can be performed.
2. Excel formulas were used to determine SNPs which were predicted to be benign by humsavar but had a phenotype identified by literature (see Humsavar\_Misidenitifed\_nsSNPs.xlsx file).
3. Genes containing pathogenic SNPs were randomized and a sample of genes which contained 1080 SNPs were chosen. Each SNP identified in GESPA was appraised to see if the predicted phenotype agreed with that identified by humsavar (see Phenotype\_Prediction\_Sample.xlsx).