GESPA User Manual

SUNY Upstate Medical University

12/1/2013

Contents

GESPA Overview

GESPA (GEnomic Single nucleotide Polymorphism Analyzer) is a program designed to quickly and accurately classify SNPs (Single Nucleotide Polymorphisms) on coding segments of the human genome by predicting both their pathogenicity and phenotype while providing several useful annotations related to the SNP. GESPA integrates data from several sources including NCBI and UCSC to make predictions based on the characteristics of specific SNPs as well as provide relevant annotations and links to specific webpages.

Predicting Pathogenicity and Phenotype of SNPs

One of the key features of GESPA is predicting the pathogenicity and phenotype of a given SNP. GESPA uses custom generated alignments of orthologous and paralogous genes to calculate a Weighted Protein Conservation Score (WPC), a measure of the frequency of a specific allele found in a SNP in the wider selection of orthologous and paralogous genes as well as a Position Specific Independent Counts(PSIC) score (Heinkoff et. Al). GESPA also obtains information related to SNPs identified in publications in the region or "hotspot" in which a SNP is found. Using these 3 main predictors, GESA is able to predict the pathogenicity of a SNP with an 87.89% balanced accuracy. This represents a 10-18% improvement over existing software. SNPs predicted to be Benign (h) are predicted to be benign due to no SNPs being identified in any publication in their "hotspot." SNPs predicted to be Benign (a) are predicted to benign due to a low PSIC and/or WPC score. Finally, SNPs predicted to be Pathogenic have a high WPC score.

GESPA predicts phenotype based on verified publications in the hotspot surrounding a SNP. GESPA looks at SNPs with a phenotype predicted by publications in a user-specified region near a SNP. (By default, GESPA looks at the region within 100 amino acids of a SNP). Once all SNPs in the region have been accounted for, GESPA predicts the phenotype to be the most frequently occurring phenotype in the region. In the event that the SNP entered had been published, the phenotype of the SNP is predicted to be the same as that predicted by the publication. Using this method, GESPA was found to have 96% accuracy in predicting the phenotype of a pathogenic SNP.

Cloud Integration

Cloud integration is crucial to the functioning of GESPA. In order to quickly provide data and represent a constantly updating body of information, an SQL-based cloud server was chosen to function as a framework for GESPA. GESPA normally collects data by mining it from various websites. After the data has been mined, it is saved to the SQL server, allowing it to be instantly accessible on a global scale. Almost the entire protein coding sequence of genes is currently available on GESPA and in the event that a gene is not present, once a user runs it and GESPA mines the data and saves it, the information will instantly be available in the future. GESPA also has features built in which allow for the cloud interface to be circumvented in order to mine more up to date information, if available. Once this up to date data is obtained, it also is accessible globally.

Annotations

GESPA provides instant access to direct and relevant annotations for each SNP. Separate and numbered nucleotide sequences, protein sequences, paralogous nucleotide alignments, paralogous protein alignments, orthologous nucleotides alignments, and orthologous protein alignments are available. For each SNP, in addition to its predictions of pathogenicity and phenotype, GESA can provide the protein change and location as well as the DNA change and location, even if only 1 value was entered by the user. GESPA provides information relating to the number of publications available for a specific SNP and can open up the related NCBI ClinVar page, which contains additional information related to a SNP including direct links to any publications on PubMed. In order to view a SNP in its larger context, GESPA is able to open up the location of a SNP in the UCSC Genome Browser allowing for seamless integration with the hundreds of annotations for genes and SNPs provided by UCSC.

GESPA also provides a host of information that was used to make the prediction of pathogenicity. A breakdown of the number and percentage of amino acids and nucleotides conserved in orthologs and paralogs as well as the specific nucleotides/ amino acids which differ in a gene from the SNP selected allows the user to see how the WPC was calculated. In addition, the PSIC score allows the user to assess the impact the prediction of pathogenicity. Although not used to predict pathogenicity, the substitution score from the BLOSUM 62 matrix is also provided.

Installation and Updates

System Requirements

Operating System:

Windows: XP SP3 or later, Vista SP2, Windows 7, Windows 8

Mac OS X 10.7.3 (Lion) or later

Hardware:

Processor: Single core processor 1 GHZ+. Quad core processor 2GHZ+ recommended for efficient parallel processing of data.

RAM: 500 MB required, 4 GB or higher strongly recommended for large batch files

Hard Drive Space: 15 Mb required

Java:

Java 1.7.0 or higher. 64 bit java recommended for running large batch files. (http://www.java.com/)

Regular Installation

1. Check to insure that your system complies with operating system, hardware, and Java requirements. Check to make sure your system has a valid internet connection.
2. Verify that Java version 1.7.0 or higher is installed on your system by visiting http://www.java.com/en/download/installed.jsp.
3. Download GESPA.rar from https://sourceforge.net/p/gespa/
4. Unzip and extract GESPA into a folder.
5. There should be folders called lib, src, and IMPORTANT_LICENSING_INFORMATION, Test_Data, and a SNPAnalyzer.jar file.
6. Ensure that the following files are in the lib folder:
   a. biosql-1.8.2.jar
   b. bytecode-1.8.2.jar
   c. commons-codec-1.7.jar
   d. commons-collections-3.2.1
   e. commons-io-2.4
   f. cmmons-lang3-3.1
   g. commons-logging-1.1.1
   h. core-1.8.2
   i. cssparser-0.9.9
   j. htmlunit-2.12
   k. httpcore-4.2.2
   l. httpmime-4.2.3
   m. jetty-http-8.1.9.v20130131
   n. jetty-io-8.1.9.v20130131
   o. jetty-util-8.1.9.v20130131
   p. jetty-websocket-8.1.9.v20130131
   q. nekohtml-1.9.18
   r. sac-1.3

   s. sequencing-1.8.2 (1)
   t. serializer-2.7.1
   u. sqljdbc4
   v. xalan-2.7.1
   w. xercesImpl-2.10.0
   x. xml-apis-1.4.01

7. If any files are missing, re-download GESPA and delete existing GESPA files.
8. Open folder called IMPORTANT_LICENSING_INFORMATION. Open and read GESPA GPL-3.0 license, HTMLUNIT License and Microsoft JDBC license. If you do not agree with any information in these licenses, please delete GESPA and its files.
9. Open file SNPAnalyzer.jar to open GESPA. Ensure that the program information panel displays that GESPA is connected to the internet, cloud, and is updated. Also check to make sure the program information panel correctly identifies the Java version installed. See troubleshooting for further information.

Installing from the Source

The following instructions use the NetBeans IDE available at https://netbeans.org/downloads/

1. Perform steps 1-6 of the regular installation.
2. Ensure that the following files are located in the src folder:
   a. Batch.java
   b. BatchManager.java
   c. ConservationPanel.form
   d. ConservationPanel.java
   e. Framework.java
   f. Gene.java
   g. GenePanel.form
   h. GenePanel.java
   i. HomologFinder.java
   j. JFrameMain.form
   k. JFrameMain.java
   l. NCBIFinder.java
   m. ReportTab.form
   n. ReportTab.java
   o. SNP.java
   p. URLReader.java
3. If any files are missing, re-download GESPA and delete existing GESPA files.
4. In the NetBeans IDE, select file→new project.
5. In the popup dialog, under categories select Java and under projects select Java Project with Existing Sources. Press Next.
6. Title your project and choose a save location. Press Next.

7. Next to the source package folders label, press the Add Folder button and select the location of the src folder in the file chooser. Press Next.
8. Check that all of the files from step 2 are included in the Included Files panel. Press finish.
9. On the right hand side of the IDE select the projects tab and expand the project that was just created.
10. Right click on the Libraries folder under the project and press Add JAR/Folder.
11. In the file chooser, open the lib folder. Select all of the jar files (Control/command-A) and press open.
12. Save the project (control/command-shift-s).
13. Press the play button on the top toolbar of the IDE to run the project. If prompted to select a main class, choose JFrameMain.java. The GESPA application should open normally to the start page. Close the application.
14. Press the hammer and broom (clean and build) in the top toolbar of the IDE. The JAR file for GESPA can be found in the dist folder of the project at its save location.

Updates

GESPA updates help to ensure that new features are added and existing bugs are fixed. When GESPA receives an update, the program will display a notification. In order to update simply press the update button; this will open a link to the update download in the default browser. Download the file and replace the SNPAnalyzer.jar file with this file after it has been updated.

User Input

Entering Genes

To enter a gene, press the new gene button on the start page. Enter the HUGO symbol (important for proper functionality) of the gene in the popup dialog and press OK.

Entering SNPs

GESPA allows SNPs to be entered in a variety of formats. To enter a gene, first enter a gene and once the option is available press the add custom SNP button and type a SNP in one of the possible formats. Once entered all SNPs are converted to their protein equivalent on the main display and their DNA and protein equivalents in the conservation and full reports. For problems related to entering SNPs, please see the troubleshooting section.

Protein: Enter the amino acid at the desired location in the genes reference sequence, followed by the location followed by the protein change. Ambiguos amino Example: I1461T

dbSNP Accession Number: Accession number for SNP from dbSNP. The SNP must be on a protein coding sequence of the gene. Accession number must be for the current gene. Example: rs6746030

DNA Nucleotide Location: Location of the SNP in DNA reference sequence. For batch files, the nucleotide change is required. Example: 4723 or 4723>a.

DNA flanking sequence: DNA nucleotide sequence starting with the nucleotide of the SNP. The flanking sequence must be long enough so there are not multiple matches in a gene. For batch files, the nucleotide change is required. Example: actthctactctatc or actthctactctatc>g

Batch Files

Batch files allow multiple genes and SNPs to be entered and processed simultaneously. Batches will first have their genes verified (the current gene being verified can be observed in the program information panel). Batch files are than processed in parallel, constrained by the number of virtual cores of CPU available. Once at least one gene in a batch has been processed, a new tab will open detailing SNP and gene information for all genes that have been processed. If any errors occur, a batch error tab will open specifying the errors that GESPA encountered while running the batch.

Running Batch Files:

1. Format batch file appropriately (see example below). Batches should be saved in .txt files. The examples below can be pasted directly into a text file and run.

   Example 1:

   ```
   Gene: GFRA1
   Y85N
   Gene: RP1
   C2033Y
   R872H
   rs2293869
   S1691P
   A1670T
   Gene: RHO
   P347A
   GACTACTACACGCTCAAGCC>C
   Gene: PTCH1
   S1132Y
   876>C
   E1438D
   ```

   Example 2 (note separation of gene and SNP by tab character):

   ```
   GFRA1	Y85N
   RP1	C2033Y
   RP1	R872H
   RP1	rs2293869
   RP1	S1691P
   RP1	A1670T
   RHO	P347A
   RHO	GACTACTACACGCTCAAGCC>C
   ```

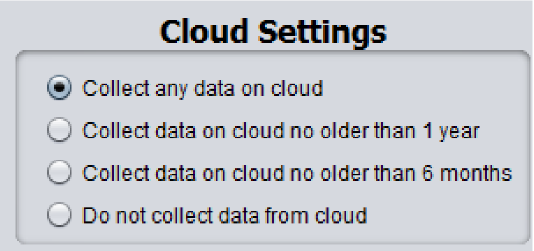
```
PTCH1   S1132Y
PTCH1   876>C
PTCH1   E1438D
```

2. In GESPA's start page, press the New Batch button and find your batch file in the file chooser at the location it was saved. Select the file and press the Add Batch button in the file chooser; the batch should start automatically.

Note: If the number of genes in a batch are greater than the maximum number of virtual cores, other genes or batches cannot be processed while a batch is running.
Note: Large batches can be very processor and RAM intensive. It is recommended that large batch files are broken up and run separately as smaller batch files.
Note: New SNPs cannot be added to a batch once it has begun.

Settings

Cloud Settings



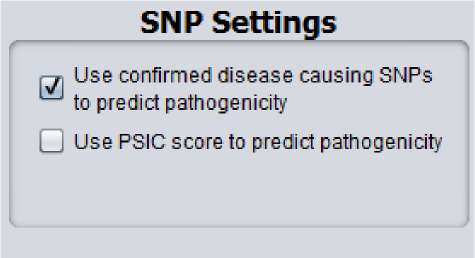
Cloud Settings

● Collect any data on cloud
○ Collect data on cloud no older than 1 year
○ Collect data on cloud no older than 6 months
○ Do not collect data from cloud

The cloud settings provide control over when data is accessed and retrieved from the cloud. By default the program collects any data available in the cloud and instantly downloads it after a gene has been entered. If newer data is desired, an interval of either 1 year, 6 months, or 0 days (Do not collect data from cloud option) can be selected.

Note: Be aware that even if new data is desired and no new data is actually available, GESPA will still mine the data from websites.
Note: Not using data available from the cloud can take significantly longer due to the variety of sources GESPA obtains data from and the creation of new alignments.

SNP Settings



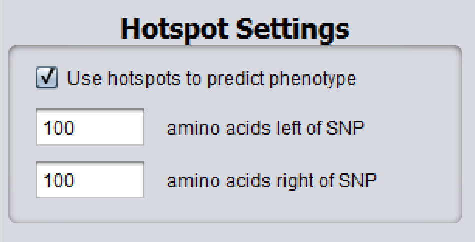
SNP Settings

☑ Use confirmed disease causing SNPs to predict pathogenicity
□ Use PSIC score to predict pathogenicity

The SNP settings provide control over which features are used to predict the pathogenicity of a SNP. By default, confirmed disease causing SNPs (i.e. SNPs on the same gene as the SNP of interest which have been confirmed pathogenic by publications) are used to predict a SNP's pathogenicity. The PSIC score is not enabled by default due to the significant computational requirements of calculating the PSIC score for a large number of SNPs, especially on genes which have a large number of orthologous and paralogous genes in their alignments. In addition the PSIC Score only provides approximately a 1% increase in specificity (0.5 % balanced accuracy increase) so should only be used on genes with few orthologous and paralogous sequences or on computers with high numerical processing capabilities.

Note: The PSIC score is not retroactively calculated (ie. Once this feature is enabled, SNPs that were already entered will not have their PSIC Score calculated if not calculated before).

Hotspot Settings



Hotspot Settings

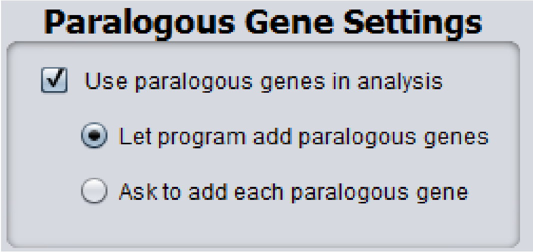
☑ Use hotspots to predict phenotype

100 amino acids left of SNP

100 amino acids right of SNP

The hotspot settings control the prediction of the phenotype of a given SNP. By default the feature is enabled; it is highly recommended that this feature is not disabled as it is not computationally expensive and provides valuable information. The default range for amino acids searched for diseases is 100 Amino acids towards 3' and 100 amino acids towards 5'. This range is fully user adjustable.

Paralogous Gene Settings



Paralogous Gene Settings

☑ Use paralogous genes in analysis

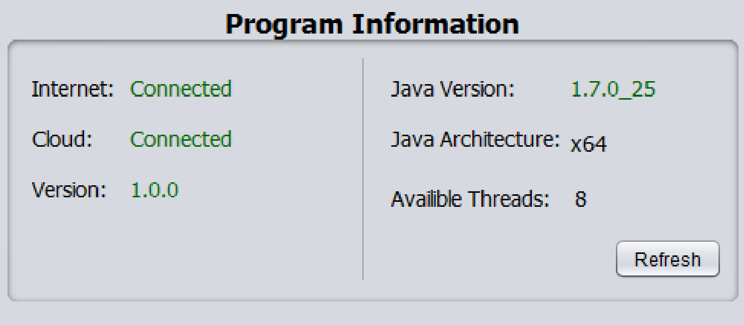
● Let program add paralogous genes

○ Ask to add each paralogous gene

The paralogous gene settings allow for control over the inclusion of paralogous genes in the analysis, if available. As a general trend, using paralogous genes provides an increase in the accuracy of pathogenicity predictions. GESPA can either automatically add paralogous genes based on its own analysis for similarity or ask the user if he or she would like to add a paralogous gene that was determined suitable by GESPA.

Note: Paralogous genes are always used, if available, when data is retrieved from the cloud.
Note: Retrieving data from the cloud must be disabled for GESPA to ask to add paralogous genes.
Note: When GESPA asks to add paralogous genes, the eventual alignment created by the user selected genes is NOT saved to the cloud.

Program Information Panel



Program Information

Internet: Connected

Cloud: Connected

Version: 1.0.0

Java Version: 1.7.0_25

Java Architecture: x64

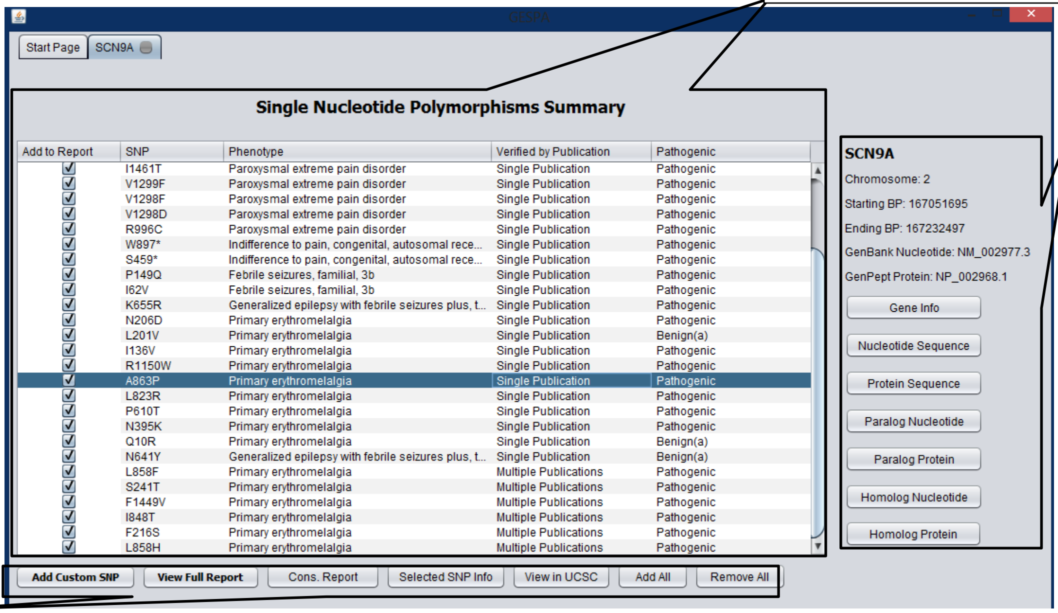
Availible Threads: 8

The program information panel provides the current status of program network connections, the version of Java in use, the type of Java being used, and the maximum number of threads usable by GESPA. For help related to the program information panel, see the troubleshooting section of the user manual. The internet and cloud labels indicate if GESPA is able to access these utilities. If GESPA is unable to access the internet, Genes or SNPs cannot be entered and if GESPA cannot access the cloud, notification of new updates cannot be received and the benefits of the cloud will not be available. In the event that an update is available, the version will be displayed in red text and the update button can be pressed. The Java version indicates the version of Java detected by GESPA. Genes or SNPS will not run properly if Java is below version 1.7.0. The Java architecture displays if java is configured to run in a 32 bit (x86) configuration of a 64 bit (x64). 64 bit Java allows for more genes to be viewed simultaneously due to a much higher memory limit provided that there is physically enough memory available for GESPA. The available threads label indicates the number of virtual cores on which GESPA can perform parallel tasks. For example, 8 available threads mean that 8 genes can be processed simultaneously. For information on changing this limit (based on number of virtual cores available to GESPA) see modifying GESPA. This operation is recommended for advanced computer users with experience in the Java programming language and may adversely impact the performance of GESPA.

Note: to save network bandwidth and computational resources, the program information panel does not update automatically and must be manually refreshed.

Viewing Annotations

Gene/Batch Summary Panel



Single Nucleotide Polymorphisms Summary

Add to Report	SNP	Phenotype	Verified by Publication	Pathogenic
☑	I1461T	Paroxysmal extreme pain disorder	Single Publication	Pathogenic
☑	V1299F	Paroxysmal extreme pain disorder	Single Publication	Pathogenic
☑	V1298F	Paroxysmal extreme pain disorder	Single Publication	Pathogenic
☑	V1298D	Paroxysmal extreme pain disorder	Single Publication	Pathogenic
☑	R996C	Paroxysmal extreme pain disorder	Single Publication	Pathogenic
☑	W897*	Indifference to pain, congenital, autosomal rece...	Single Publication	Pathogenic
☑	S459*	Indifference to pain, congenital, autosomal rece...	Single Publication	Pathogenic
☑	P149Q	Febrile seizures, familial, 3b	Single Publication	Pathogenic
☑	I62V	Febrile seizures, familial, 3b	Single Publication	Pathogenic
☑	K655R	Generalized epilepsy with febrile seizures plus, t...	Single Publication	Pathogenic
☑	N206D	Primary erythromelalgia	Single Publication	Pathogenic
☑	L201V	Primary erythromelalgia	Single Publication	Benign(a)
☑	I136V	Primary erythromelalgia	Single Publication	Pathogenic
☑	R1150W	Primary erythromelalgia	Single Publication	Pathogenic
☑	A863P	Primary erythromelalgia	Single Publication	Pathogenic
☑	L823R	Primary erythromelalgia	Single Publication	Pathogenic
☑	P610T	Primary erythromelalgia	Single Publication	Pathogenic
☑	N395K	Primary erythromelalgia	Single Publication	Pathogenic
☑	Q10R	Primary erythromelalgia	Single Publication	Benign(a)
☑	N641Y	Generalized epilepsy with febrile seizures plus, t...	Single Publication	Benign(a)
☑	L858F	Primary erythromelalgia	Multiple Publications	Pathogenic
☑	S241T	Primary erythromelalgia	Multiple Publications	Pathogenic
☑	F1449V	Primary erythromelalgia	Multiple Publications	Pathogenic
☑	I848T	Primary erythromelalgia	Multiple Publications	Pathogenic
☑	F216S	Primary erythromelalgia	Multiple Publications	Pathogenic
☑	L858H	Primary erythromelalgia	Multiple Publications	Pathogenic

SCN9A
Chromosome: 2
Starting BP: 167051695
Ending BP: 167232497
GenBank Nucleotide: NM_002977.3
GenPept Protein: NP_002968.1

The main table provides important information for each SNP: amino acid location, phenotype, the number of related publication, GESPA's predicted pathogenicity. The phenotype is either GESPA's prediction (preceded by Possibly) or a phenotype identified by a publication.

The lower portion of the panel allows access to functions related to SNPs such as adding SNPs, viewing reports, and viewing SNPs in UCSC and NCBI.

The right portion of the panel provides information specific to a gene: viewing the gene on NCBI and nucleotide and protein sequences and alignments.

Sequences and Alignments

GESPA offers nucleotide and amino acid sequences for each gene as well as nucleotide and amino acid paralogous and orthologous alignments when available. To access this information, simply press the buttons corresponding to the information desired on the right side of a gene's main information tab. To access this information in a batch file select a SNP located on the gene of interest on the batch summary tab. All available gene information can then be accessed as normal.

Note: nucleotide alignments are colored by nucleotide and protein alignments are colored by amino acid family (DE-RED, Acidic. AVFPMILW- Small, ORANGE. RK-Basic, BLUE, STYHCNGQ – Hydroxyl + sulfhydryl + amine + G, GREEN)

Viewing SNPs in the UCSC Genome Browser

GESPA allows for individual SNPs to be viewed in the UCSC genome browser in context of a portion of their gene.

1. Select a SNP by highlighting it in a gene/batch main panel and press the View in UCSC button at the bottom of the interface. The SNP and nearby nucleotides will open in a UCSC genome browser page in the default browser.
2. To locate the SNP itself in the genome browser, scroll down to the variations and repeats section and expand it.
3. Under the flagged SNPs drop down menu, select full.
4. Click the refresh button. In the browser interface clinically flagged SNPs will appear in red.
5. If the SNP has a dbSNP accession number, it can be found on either the extreme right or left sides of the genome browser interface. Click on the flagged SNP for additional annotations.

Viewing SNP and Gene Information in NCBI

GESPA supports viewing information for SNPs (when available) as well as genes using NCBI. This information includes relevant publications for genes and SNPs as well as links to additional databases.

Gene Information

To view gene information, click Gene Info at the right side of a Gene/Batch Summary Panel. An NCBI page displaying information related to the gene should appear. The webpage should allow for navigation to external and internal databases and display additional relevant information for the gene, including any relevant publications.

SNP Information

To view SNP information, select a SNP by highlighting it in a Gene/Batch Summary panel and press the Selected SNP Info button at the bottom of the interface. If a page is available on NCBI for the selected SNP, the page will open on the default web browser. The page should contain annotations for the SNP and related links to external and internal databases. To view publications for the SNP, press the Evidence tab in the lower portion of the interface, if available.

Note: The nucleotide locations provided for SNPs may differ from GESPAs nucleotide locations due to different nucleotide sequences being used. GESPA displays the accession numbers for the nucleotide and protein sequences it uses in the Gene/Batch Summary Panel

The Conservation and Full Reports

GESPA's conservation and full reports show breakdowns of both protein and nucleotide conservation. To access the reports, select the desired SNPs by checking their add to report boxes in the Gene/ Batch Summary Panel. Then press the corresponding button for the desired report located at the bottom of the Gene Summary Panel.

Note: By default, all custom SNPs added are selected to be added to the reports. In batch files, SNPs associated with publications are not added by default while these SNPs are added by default when running single genes.

Troubleshooting

If problems cannot be resolved using the information available below, post a question on the GESPA discussion forums at http://sourceforge.net/p/gespa/discussion/

Connectivity and Program Information Panel

Internet status displaying "Not connected"

1. First verify to make sure that your system is actually connected to the internet by navigating to http://www.ncbi.nlm.nih.gov/ in the default browser.
2. Try connecting GESPA to a different network (example guest network, home network) to see if the problem persists.
3. Check to make sure no operating system or antivirus firewall is blocking access to the program. Make sure that Java is allowed to access the internet.
4. Check with your network administrator to see if certain applications are blocked from accessing the internet.

Cloud Status Displaying Connection Error

1. Try connecting GESPA to a different network (example guest network, home network) to see if the problem persists.
2. Verify that communications on port 1433 are not blocked by either the operating system or antivirus software. Port 1433 is used to communicate to SQL-based servers. Visit the following pages for more information:
   http://windows.microsoft.com/en-us/windows/open-port-windows-firewall#1TC=windows-7
   http://support.microsoft.com/kb/287932
3. Check with your network administrator to see if port 1433 is blocked from communications.

Incorrect Java Version or Architecture

Uninstall Java completely from your system. Check to make sure it is uninstalled by visiting the link http://www.java.com/en/download/installed.jsp . Reinstall the appropriate version of Java and rerun GESPA.

SNP Input

dbSNP Input Not Accepted

GESPA only accepts dbSNP accession numbers for SNPs on protein coding segments of the gene. Verify that the accession number entered corresponds to a SNP on a protein coding region. At times, protein changes on dbSNP are identified by multiple different protein sequences and GESPA cannot identify the amino acid locations corresponding with the sequence which it is using. In this case, enter the SNP in amino acid format instead of as a dbSNP accession number.

Nucleotide at Location Entered is on Intron

When copying a nucleotide sequence into GESPA for entry as a custom SNP, ensure that the entire sequence is located only on an exon. Also, be aware that the nucleotide sequence provided by GESPA contains some areas on introns.

Modifying GESPA (Advanced)

GESPA is an open source program and as such is modifiable. To modify GESPA, follow the instructions for installing from the source and before the final step, modify the source code as desired. It is highly recommended that GESPA's data retrieving methods are not modified due to their complex connected methods which could render the program unusable. Sample modifications follow below.

Changing Maximum Number of Threads

In NetBeans, open the GESPA project and the JFrameMain.java file. Locate the static final variable MAX_CONCURRENT_THREADS. The current value is based on the number of virtual cores available to Java and can be adjusted as desired. Be aware that increasing the number past the number of virtual cores can cause performance issues if the system is incapable of handling the processing.

Changing Cutpoint Values in the Pathogenicity Prediction Algorithim

GESPA's default pathogenicity prediction algorithm is optimized for the highest possible balanced accuracy. If a higher sensitivity or specificity is desired, the cutpoints of the algorithm can be adjusted.

1. Download the testData.xlsx file to experiment how changing the PSIC or WPC cutpoints can affect sensitivity, specificity, and balanced accuracy.
2. In NetBeans, open the GESPA project and the Framework.java file. Locate the method private String predictPathogenic(double protCons, double PSICScore, int substScore){}
3. Adjust the cutpoints as desired for the PSICScore and protCons (WPC Score).

Note: If !JFrameMain.useLit (program doesn't use SNPs identified in literature to predict pathogenicity) different cutpoints are used than if JFrameMain.useLit==true. By default, JFrameMain.useLit==true.

Note: If redistributing modified or unmodified versions of GESPA, all conditions outlined in GESPA's GPL 3.0 License must be adhered to