GeneTerpret Documentation

Dr. Mohsen Hosseini, Dr. Roozbeh Manshaei, Veronica Andric, Sean DeLong, Esha Joshi, Priya Dhir

December 2019

Contents

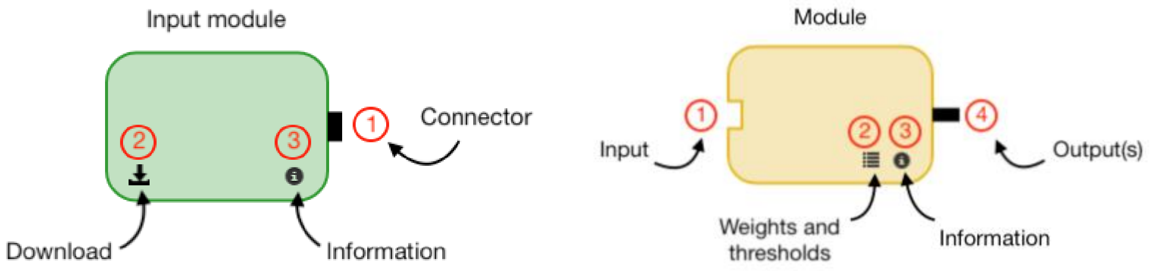
1 Introduction

GeneTerpret is a bioinformatics tool used to assess whole-genome and whole-exome sequencing data, identify and interpret the significance of genomic variants. GeneTerpret allows users to determine gene-disease relationships and variant pathogenicity by collecting variant forms of evidence using automated modules. Users can also visualize the distribution of variants in their data based on their clinical significance.

The software contains 10 modules: ExPhenosion, CanGene modules: gene expression; cross-species: mouse; cross-species: zebrafish; protein-protein interactions; homology, KING (Known Involved Genes), the Variant Interpretation program, Gene Validity and Causality.

2 Overview of Interface

2.1 General Usage



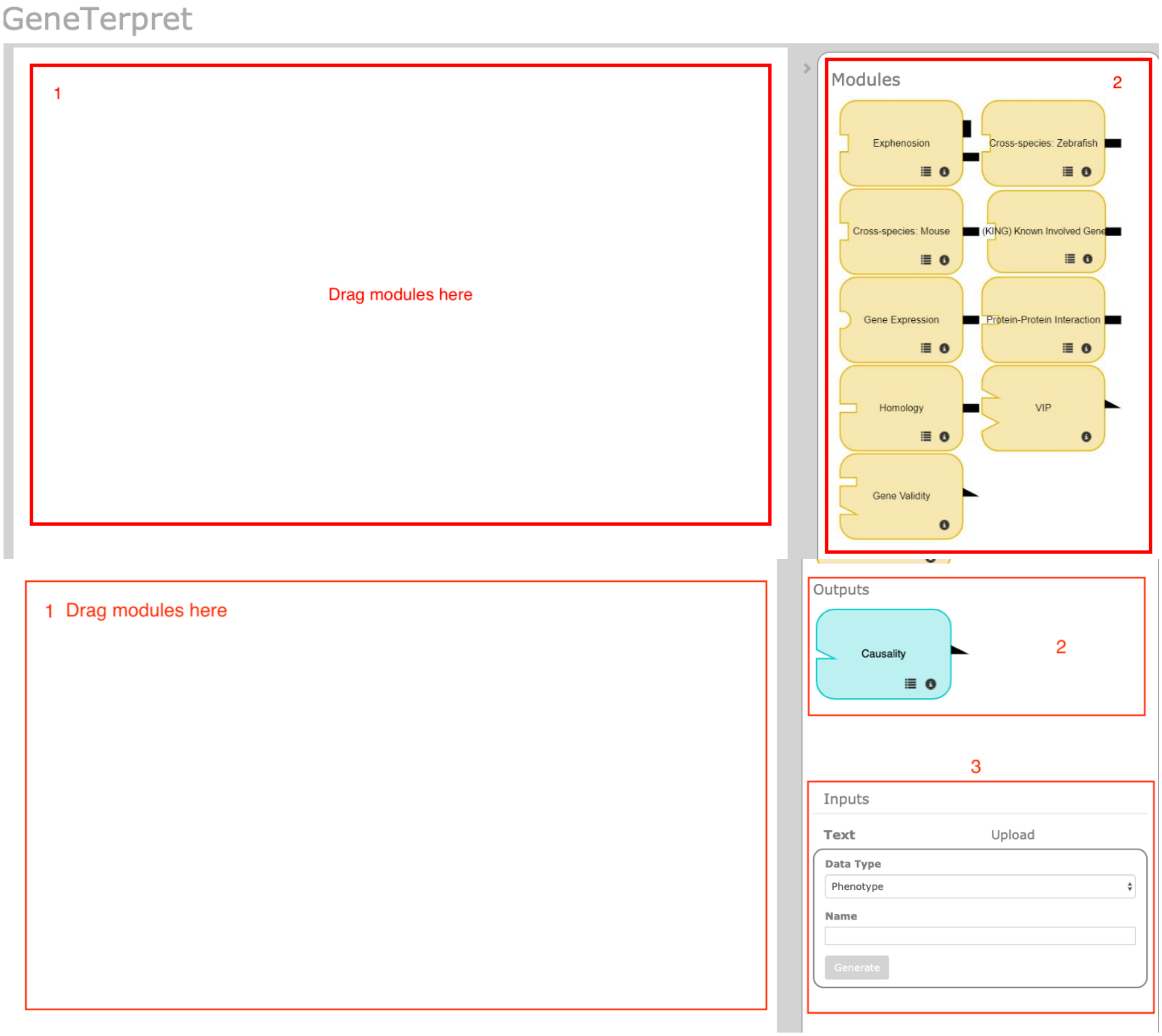
There are two types of modules: input modules to input phenotypes, genes, or VCF files, and modules to collect and generate evidence to find candidate genes, establish clinical validity and pathogenicity based on your input(s).

Weights

__Weight:__ numerical value from 1-5, adjustable by user. Increasing value indicates this module being weighed heavier when quantifying the strength of evidence for a gene-disease relationship in the Gene Validity module. E.g. if the ExPhenosion module is given a weight of 5, variants in a user-inputted VCF file that are identified in the genes outputted from ExPhenosion will be considered as variants with high clinical validity.

2.2 Overview

When GeneTerpret is opened, the interface will show a blank canvas where the various modules shown on the right (detailed in later sections) can be dragged to and interconnected. In the following sections is a tutorial detailing how to use GeneTerpret.



Basics:

1. Drag modules to the canvas to interconnect the inputs and outputs

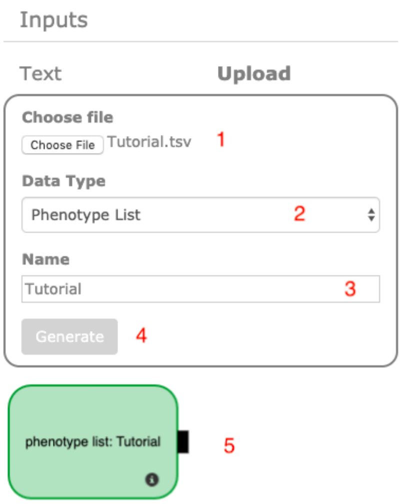
2. Scroll down to see the Causality output module to visualize the distribution of your variants based on pathogenicity (*details in 3.8*)

3. Input phenotype and gene(s) in the form of text, or upload phenotypes and genes in the form of tab delimited files, in addition to VCF and PED files

2.3 Tutorial

Below is a simple example of how to use GeneTerpret, upload files, connect modules, and specifically how to use Exphenosion, Gene Validity and VIP.

A. Creating your inputs



Inputs

Text Upload

Choose file

Choose File Tutorial.tsv

Data Type

Phenotype List

Name

Tutorial

Generate

1. You can either create your inputs using text or uploading files. Here we are uploading a tab delimited file called *Tutorial.txt* of phenotypes.

Note: try using our ExPhenosion example input on [*www.geneterpret.com*](http://www.geneterpret.com)*. Some modules require the phenotype list to contain Mondo IDs for the phenotypes or diseases of interest, these can be found at* [*https://www.ebi.ac.uk/ols/ontologies/mondo*](https://www.ebi.ac.uk/ols/ontologies/mondo)*.*

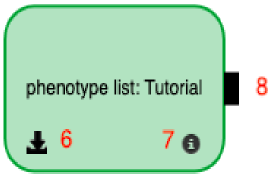
2. Specify what type of file you are uploading: VCF file, PED file, gene list or phenotype list. Here we specify ‘phenotype list’.

3. Name your input node, here we name it ‘Tutorial’.

4. Click generate to create the input node. You should see a green module show up below.

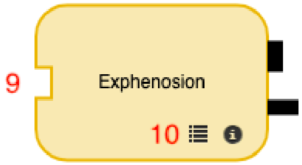
5. You can now drag this module to the canvas and connect it to other modules.

B. Connecting modules



6. After you have dragged your input node to the canvas, a download symbol will appear. This indicates your file has been uploaded successfully. You may click this button to download your input file.

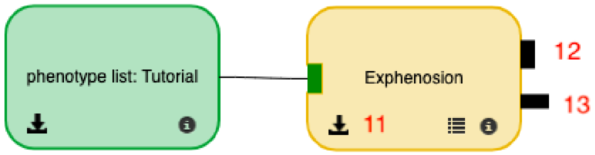
7. Click the ⓘ symbol on all modules to get information and a description of what the module does.



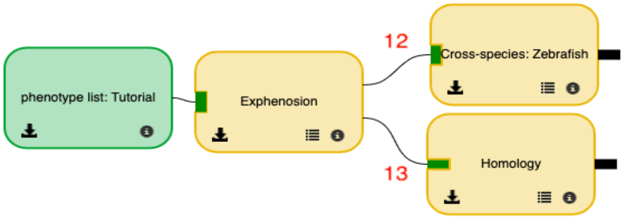
8. This rectangle connector is used to connect the input node to other modules, such as ExPhenosion. Drag the ExPhenosion module onto the canvas. All connectors have a corresponding matching shape-matching input connection.

9. The rectangle connector from your input node can be dragged and attached onto the ExPhenosion module.

10. Click this symbol to change the weights for this module.



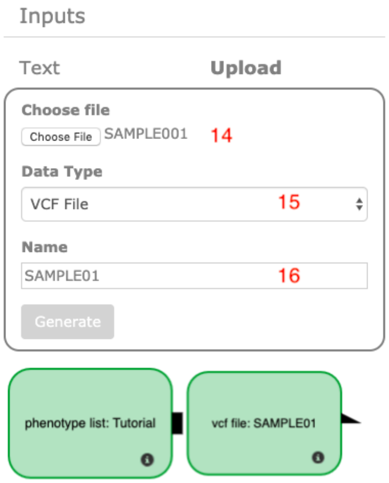
11. When you connect your input node and module, GeneTerpret will run and once the output is ready, the download button will appear. In this case, our output will create a *.zip* file with two *.txt* files: a gene list and phenotype list.



12. This is the connector for the output of the phenotype list generated by ExPhenosion. To continue collecting more evidence, you may connect this to other modules such as Cross species: Zebrafish.

13. This is the connector for the output of the gene list generated by ExPhenosion. To continue collecting evidence, you may connect this to other modules like Gene Validity.

C. Using Gene Validity and the Variant Interpretation Program


Inputs

Text Upload

Choose file

Choose File SAMPLE001 14

Data Type

VCF File 15

Name

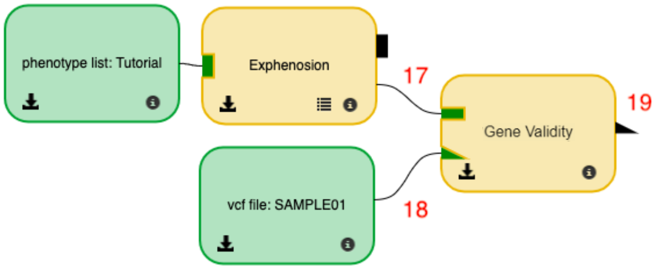
SAMPLE01 16

Generate

14. To use Gene Validity and VIP, you will need to upload a VCF file. It can be an individual VCF, trio-VCF, or cohort VCF. *Note: try using our sample VCF files on www.geneterpret.com.*

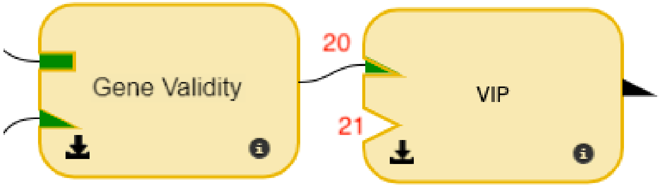
15. Specify that the file type is a VCF.

16. Name your VCF file, here we are using our individual VCF file, available for download at www.geneterpret.com. Click Generate, and your VCF node will be ready to drag onto the canvas.



17. For Gene Validity, you will need to connect a gene list to the module. In this case, we are using the gene list generated from ExPhenosion in **step 13**. You may also upload your own gene list or connect gene list outputs from other modules.

18. You will also need to connect a VCF file. Here, we are using the individual VCF we uploaded.



19. At this point, you can download the new VCF file Gene Validity generates to see the ranking of variants based on evidence supporting the gene-disease relationship or connect this file to VIP to generate classifications.

20. Connect the output of Gene Validity to VIP. Once completed, the download button will appear. The output will create a *.zip* file with three VCF file(s): *de novo variants*, *pathogenic/likely pathogenic variants, and all variants* (see 3.7 for more information).

21. If you are doing trio-based analysis, you will need a PED file with pedigree information. It can be uploaded and connected here to VIP.

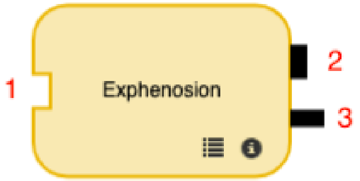
You're done the tutorial!

3 GeneTerpret Modules

3.1 ExPhenosion

The Expanded Phenotype Exploration (ExPhenosion) module can be used to obtain candidate genes related to phenotype(s). It can also be used to find an expanded list of related phenotypes for further exploration. This module uses data from HPO and MeSH databases and takes ≤1 minute to run.

Note: there is potential for redundancy in outputs when using this module in addition to the KING module as both use the same databases within their searches for finding associated genes for a particular phenotype. We recommend using the KING module to find strong evidence of associated genes for a particular phenotype.



1. Input: input the full term name of a phenotype or its HPO ID (eg., "HP:0001636" using text) or phenotype list (uploaded in a .tsv or .txt file). The format of this file should be a list of these terms or IDs separated by newlines, with the first line being "phenotype". HPO terms can be found at https://hpo.jax.org/app/
2. Phenotype list output: provides expanded phenotype list in addition to those in the input
3. Gene list output: provides a list of candidate genes. This output can be connected to other modules like Cross-species: mouse and zebrafish, protein-protein interactions, or Gene Validity.

3.2 CanGene Modules

The CanGene modules are a group of modules that provide candidate gene predictions through various forms of evidence:

i. Protein-protein interactions: determined through protein-protein interactions (PPI) or homology

ii. Cross-species: based on similar genes found in animal models such as mouse and zebrafish with similar phenotypes

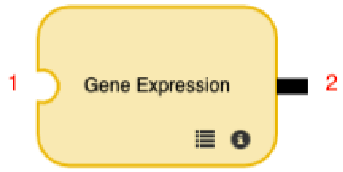
iii. Gene expression: based on orthologues genes causing similar phenotypes in model organisms such as zebrafish and mice

iv. Homology: based on homologues to the known associated genes for a particular phenotype

There are 4 CanGene modules to obtain candidate genes from using these various forms of evidence.

3.2.1 Gene Expression

The Gene Expression module will provide a list of candidate genes expressed in a particular tissue, relevant to your phenotype or disease of interest. This module uses data from the EMBL-EBI Expression Atlas database.



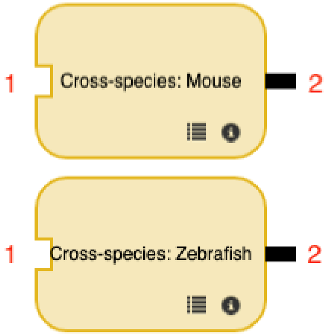
1. Input: input a tissue type (using text) or a list of tissues (uploaded in a *.tsv* or *.txt* file). In the text file, the tissues in the list should be separated by new lines with the first line being "tissue".

2. Output: provides a list of candidate genes expressed in the tissues inputted. This output can be connected to other CanGene modules like homology or protein-protein interactions, or the Gene Validity module.

Note: this module takes longer to process data – anticipate waiting a couple minutes to receive an output (i.e. for tissue "heart", gene expression takes ~5 minutes to run on a Mac).

3.2.2 Cross-species: Mouse and Zebrafish

The Cross-species: Mouse and Zebrafish modules will provide a list of candidate genes for a disease(s) expressed in mice or zebrafish, returning equivalent human orthologues. This module uses data from the MGI, Monarch and ZFIN databases. These modules take ≤2 minutes to run.

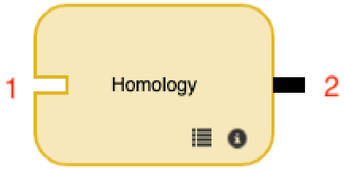


1. Input: input disease using text or upload a disease list (as a *.txt* file). The text file should have the "phenotype" header and each subsequent line in the text file should be a MONDO ID, which can be found at: https://www.ebi.ac.uk/ols/ontologies/mondo. E.g., MONDO:0006664 for "atrial heart septal defect".

2. Output: provides a list of candidate genes expressed in mice and/or zebrafish. This output can be connected to other CanGene modules like homology, protein-protein interactions, or the Gene Validity module.

3.2.3 Homology

The Homology module will provide a list of candidate genes which are homologues, or genes related to each other via duplication events in other species, to a list of input genes. This module uses data from the ENSEMBL Paralogues database. This module takes ≤2 minutes to run.

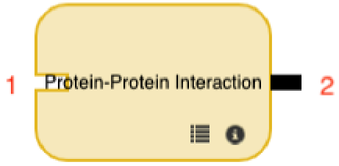


1. Input: input a gene list obtained from other CanGene modules such as Cross-species: Mouse or Zebrafish, protein-protein interaction, or the KING and ExPhenosion modules.

2. Output: provides an expanded gene list of homologues. This output can be connected to Gene Validity or other modules accepting gene lists, to collect further evidence.

3.2.4 Protein-protein Interactions

The Protein-protein Interactions module will provide a list of candidate genes that uniquely interact within given input of gene(s). This module uses data from BioGRID and takes ≤2 minutes to run.

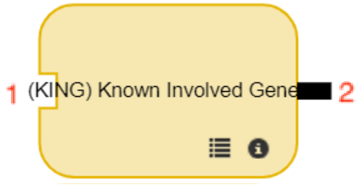


1. Input: input a gene list obtained from other CanGene modules such as Cross-species: Mouse or Zebrafish, or the KING and ExPhenosion modules.

2. Output: provides a list of genes that uniquely interact with the input genes. This output can be connected to Gene Validity or other modules accepting gene lists, to collect further evidence.

3.3 KING: Known Involved Genes

The Known Involved Genes or KING module, will find strongly associated genes with a particular phenotype(s). This module uses data from OMIM, Orphanet, Clinvar, MedGen and takes ≤1 minute to run.



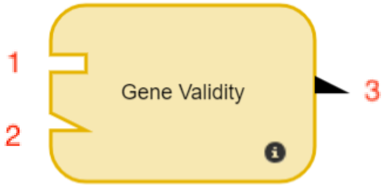
1. Input: input disease using text or upload a disease list (as a *.txt* file). The text file should have the "phenotype" header and each subsequent line in the text file should be a MONDO ID, which can be found at: https://www.ebi.ac.uk/ols/ontologies/mondo. E.g., MONDO:0006664 for "atrial heart septal defect".

2. Output: provides a list of strongly and highly associated genes with the input, drawing from various databases. This output can be connected to the CanGene modules to collect further evidence or to Gene Validity.

3.4 Gene Validity

The Gene Validity module is used to quantify the measure of the strength of evidence that supports a gene-disease relationship. It is quantified on the scale of no evidence, limited, moderate, strong, to definitive evidence. This module will consolidate the given gene lists (from CanGene modules, ExPhenosion, KING) and append a score to each gene based on frequency, user-assigned weights and thresholds. We recommend that KING output be considered as strong evidence (known genes) as it refers to genes published and directly associated with human phenotypes in four well-established databases. For all other modules, we recommend them to be treated as limited evidence (candidate genes) for further exploration. This module takes ≤1 minute to run.

1. Input for gene list: input a gene list obtained from any of the following: ExPhenosion, CanGene modules, KING.



2. Input for VCF: input an uploaded VCF file (as a *.txt* file). *Note: See 3.5.1 for required formatting of VCF files to run this module. Do not upload an .xlsx file.*

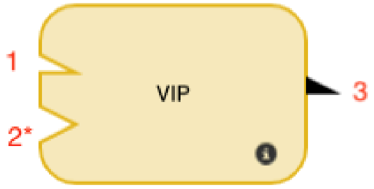
3. Output: provides a new VCF file with a new column appended with a score for each gene in the gene list and VCF. The thresholds and weights which you set for modules that obtain the gene list for will affect the validity scores produced. This output can be connected to VIP to obtain classifications for variants in the file.

It is important to note that the module parameters set by the user, such as the thresholds and weights, will impact the validity scores produced.

3.5 The Variant Interpretation Program (VIP)

The Variant Interpretation Program or VIP module establishes and appends pathogenicity to a given VCF file.

1. Input: input an uploaded VCF file (as a *.txt* file) or VCF output from Gene Validity. *Note: See 3.5.1 for required formatting of VCF files to run this module*. *Do not upload an .xlsx file.*



2. Input*: if you are using VIP to analyze a family-based VCF or trio-VCF file, you will need to input a *.PED* file for the pedigree information here. For cohort and individual analysis, it is not required. *Note: See 3.5.2 for required formatting of PED files to run this module.*

3. Output: provides three new VCFs in a *.zip* file with a new column appended with pathogenicity classifications for each variant:

i. A VCF file with only *de novo* variants (if applicable).
ii. A VCF file with only *Pathogenic* and *Likely Pathogenic* variants.
ii. A VCF file with all variants.

Information about ACMG criteria met to establish the classification is also provided: 1 = criteria met, 0 = criteria not met. This output can be connected to the Causality module to visualize the distribution of variants and lasso-select for particular variants. These classifications and criteria are based upon the ACMG's Standards and Guidelines for the Interpretation of Sequence Variants (Richards *et al.* 2015, PMID:25741868).

Note: Depending on the file size, this module may take longer to process data – anticipate waiting a couple minutes to receive an output (i.e. for a cohort VCF of 5 samples, it takes ~3 minutes to run on a Mac). Additionally, we emphasize that the variant pathogenicity classifications created from VIP **do not** conclusively indicate if a variant is pathogenic in a particular patient, nor do they signify clinical significance.

3.5.1 VCF file formatting

In order for the Gene Validity and VIP modules to be able to read your VCF files, the VCF requires functional annotation and specific headers with information about: gene-based annotation, region-based annotation, filter-based annotation (allele frequencies, *in-silico* analyses, etc). This can be achieved through ANNOVAR (http://annovar.openbioinformatics.org/).

To make sure your input VCFs are formatted correctly to be able to run on these modules, please see the headers in our example VCF datasets on http://www.geneterpret.com.

3.5.2 PED file formatting

When performing family bases analyses on trio-VCFs, a *.PED* file with pedigree information is required to establish pathogenicity classifications using VIP. A *PED* file is usually a tab-delimited file with six mandatory columns:

i. Family ID
ii. Sample ID
iii. Paternal ID/Paternity
iv. Maternal ID/Maternity
v. Sex (1=male; 2=female; other=unknown)
vi. Phenotype (-9=missing, 0=missing, 1=unaffected, 2=affected)

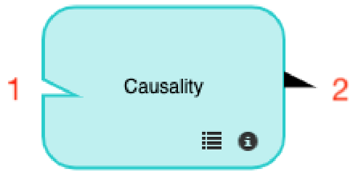
```
family_id     sample_id       father_id       mother_id       sex     phenotype
HSC_001       6711234 9145028 7234560 2       2
HSC_001       7234560 0       0       2       1
HSC_001       9145028 0       0       1       1
```

When uploading your *.PED* file, ensure these columns are included and the file is saved and uploaded with the *.ped* extension*.*

For more information, see PLINK basic usage and data formats: http://zzz.bwh.harvard.edu/plink/data.shtml.

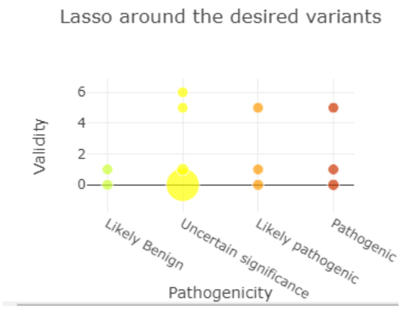
3.6 Causality

The Casuality module provides a graph and visual representation of the pathogenicity of variants and their clinical validity scores. This module requires output from Gene Validity inputted into VIP to append pathogenicity classifications.



1. Input: input the VCF output from the VIP module after running Gene Validity as well

i.e. Run Gene Validity on SAMPLE001 VCF and gene list from KING -> input result into VIP to generate classifications -> input result into Causality



2. Output:

i. Graph with distribution of variants based on their pathogenicity

ii. VCF file filtered for lasso-selected variants chosen from graph, can be downloaded using download button. The VCF file will contain a list of ranked and sorted variants based on the number of evidences extracted from the validity modules and pathogenicity terms.

4 Frequently Asked Questions

How do I know if my files have been uploaded correctly?
When you upload a file (VCF, PED, gene or phenotype list) and see a download button appear – your file has uploaded correctly. You can download it to verify it contains the correct contents.

The download button won't appear when I upload my files.
If you do not see the download button appear, please check that: (1) you are uploading the correct file type (usually *.txt* file), (2) you are not uploading a *.xlsx, .xlsm,* or *.csv.*

My two modules won't connect.
There are modules that only take inputs of a certain type (i.e. Gene Expression). Ensure that the output of one module matches the shape for input of the second module.

When downloading my output from VIP, I only have two VCF files in the folder.
Not all VCFs will have variants that are *de novo* or *Pathogenic/Likely Pathogenic*. In this case, the module won't always produce three separate files. However, you will always see classifications for all variants in one VCF, including variants from benign, uncertain significance, to pathogenic.

My outputs from modules like cross-species: mouse or cross-species zebrafish are empty.
For particularly less-studied phenotypes and genes, there will not be information on paralogues in animal models such as mouse or zebrafish. This is also true for other modules like KING and homology where there might not be established data linking phenotypes to genes.

5 Contact Us

If you encounter any problems when running or using GeneTerpret, please contact us on our website at www.geneterpret.com or email at us geneterpret@gmail.com.

6 References

Databases used

ExPhenosion

The Human Phenotype Ontology (HPO): Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Research. (2018) doi: 10.1093/nar/gky1105. http://www.human-phenotype-ontology.org

MeSH: Medical Subject Headings [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2019 December 18]. Available from: https://www.ncbi.nlm.nih.gov/MeSH

CanGene

EBI Expression Atlas: Irene Paptheodorou, Nuno A. Fonseca, Maria Keays, Y Amy Tang et al. Expressional Atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Research. (2018) doi: 10.1092/nar/gkx1158. https://www.ebi.ac.uk/gxa/home

MGI Mouse Genome Informatics: Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE. The Mouse Genome Database Group. Mouse Genome Database (MGD) 2019. Nucleic Acids Research. (2019) doi: 10.1093/nar/gky1056. http://www.informatics.jax.org/

Monarch Disease Ontology: Christopher J. Mungall, Julie A. McMurry, Sebastian Kohler, James P. Balhoff et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. (2017) doi: 10.1093/nar/gkw1128. https://monarchinitiative.org/

The Zebrafish Information Network (ZFIN): Ruzicka L, Howe DG, Ramachandran S, Toro S, et al. The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources. Nucleic Acids Research. (2019) doi: 10.1093/nar/gky1090. https://zfin.org/

ENSEMBL: Sarah E Hunt, William McLaren, Laurent Gil, Anja Thormann, et all. Ensembl variation resources. Database. (2018). doi: 10.1093/database/bay119. https://useast.ensembl.org/index.html

BioGRID: Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Research. (2006). doi: 10.1093/nar/gkj109. https://thebiogrid.org/

Known Involved Genes (KING)

Online Mendelian Inheritance in Man (OMIM): Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, Victor A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research. (2005). doi: 10.103/nar/gki033. https://www.omim.org/

Orphanet: Orphanet: an online database of rare diseases and orphan drugs. Copyright, INSERM 1997. Available at http://www.orpha.net.

ClinVar: Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang et al. ClinVar: a public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research. (2014). doi: 10.1093/nar/gkt1113. https://www.ncbi.nlm.nih.gov/clinvar/

MedGen: MedGen [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2019 December 18]. Available from: https://www.ncbi.nlm.nih.gov/medgen/