PhenUMA Tutorial

Contents

INTRODUCTION

What is PhenUMA?

PhenUMA is a web application for the integration and visualization of biomedical networks. These networks are based on phenotypic and functional relationships (interactions) retrieved from different types of inputs queried by the user as gene IDs, OMIM genes, OMIM diseases, Orphan diseases or phenotypes.

Therefore, it provides a framework enabling the integration of known and emergent interactions between elements using different data sources. PhenUMA works with two types of interactions, named as known or inferred, can be retrieved using PhenUMA (for more info see: Where the information comes from?).

The known interactions come directly from information or relational databases (for more info see Known relationships section):

Genetic association studies (OMIM or Orphanet). For instance, genes associated with diseases as well as diseases associated with genes.
Physical interactions (STRING)
Metabolic interactions (base on positive correlation fluxes calculated from metabolic network Recon 1 Veearami et al. 2009)

The inferred interactions are deduced from the known relationships or calculated from databases of other biocomputational resources:

Phenotypic semantic similarities based on Human Phenotype Ontology (HPO). Functional semantic similarities based on Gene Ontology (GO).
Disease co-associations by at least one shared gene.
Gene co-associations by at least one shared disease.

PhenUMA is recommended for:

Clinical or biomedical researchers interested to study the underlying pathological and biological relationships in their experimental datasets. These datasets have to be composed of genes, diseases or clinical features (phenotypes).

PhenUMA is NOT recommended for:

Clinical diagnosis.

Where the information comes from?

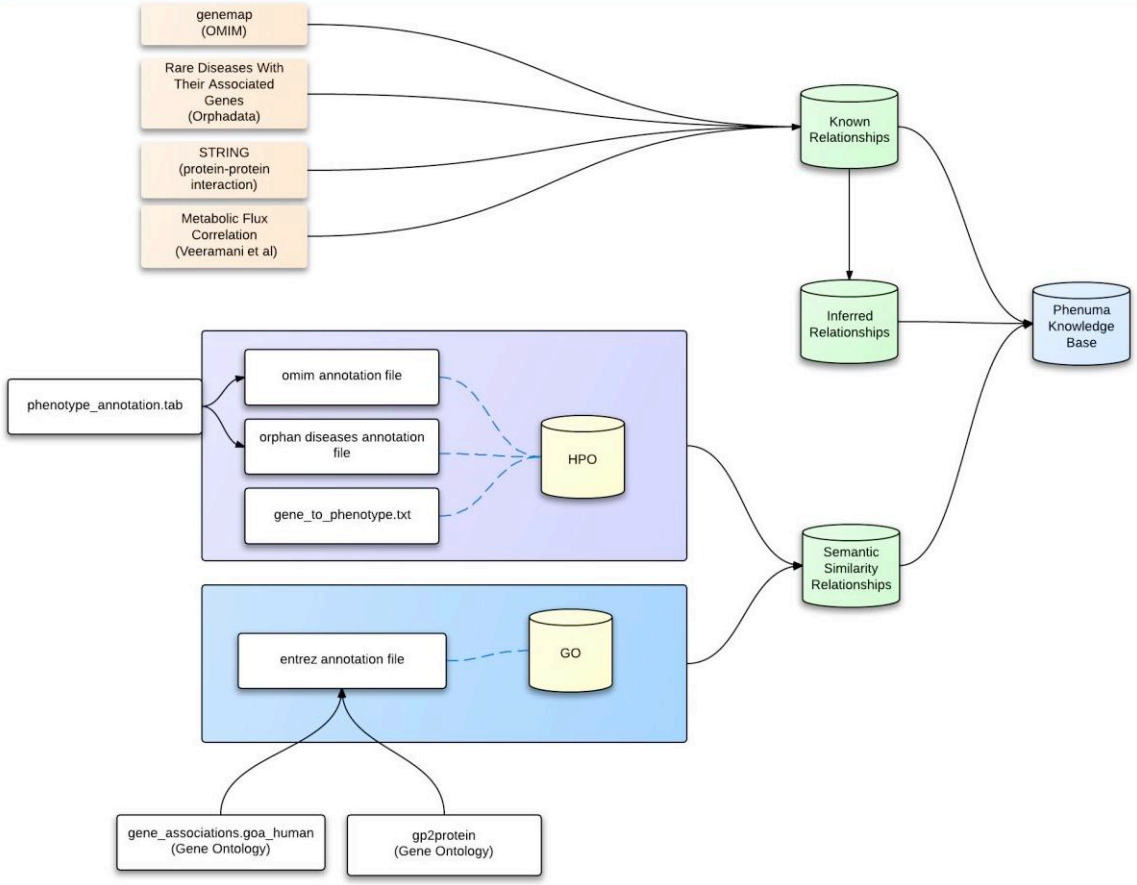


Figure 1 This schema represents the sources of the information included in the knowledge base. The arrows indicate that the source node has been used to generate the target node (for example the genemap file has been used to create the known relationships section). The dashed lines indicate that two files have been used together (annotation file – ontology).

1. Known Relationships

OMIM to gene

The relationships between genes and OMIM disorders have been extracted from the files: *genempa.txt* and mim2gene.txt provided by OMIM.

Orphan disease to gene

Rare Diseases With Their Associated Genes (http://www.orphadata.org/cgi-bin/inc/product1.inc.php) is a XML file, which is included in www.orphadata.org. *We* used this file to include in our database the relationships between rare diseases and genes.

Metabolic

At the beginning, we are using those positive metabolic flux correlations between enzyme coding genes from the version of Recon 1 analysed by Veeramani, B., & Bader, J. S. (2009). In a future we are considering to include a addtional metabolic interactions using alternative flux coupling criteria.

Veeramani, B., & Bader, J. S. (2009). Metabolic Flux Correlations, Genetic Interactions, and Disease. Journal of Computational Biology, 16(2), 291-302.

Protein-Protein Interaction

The protein-protein interaction relationships have been included from STRING dataset.

2. Inferred Relationships

The inferred relationships are calculated using the known relationships extracted from OMIM and Orphadata. This type of links relates:

Genes sharing OMIM diseases
Genes sharing Orphan diseases
OMIM diseases sharing genes
Orphan diseases sharing genes

The inferred networks are calculated using the complete bipartite networks of known relationships. As mentioned above, one of these networks includes relationships among OMIM diseases and genes and the other one does it among orphan diseases and genes. In the Figures 2 and 3 is shown how the inferred relationships are deduced:

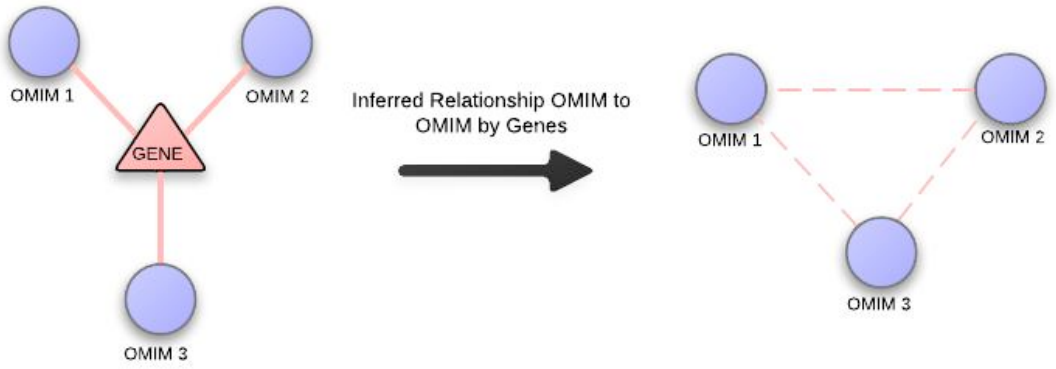


Figure 2 In the network on the left side of the figure there is a gene related with three OMIM diseases. The outcome inferred network links all the diseases among them. The associated score to each relationship is 1, since each pair of diseases shares only one gene.

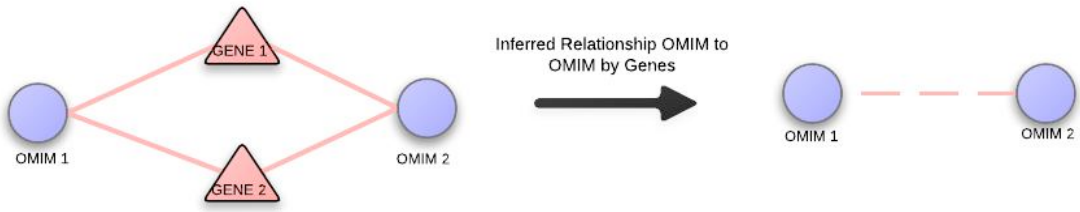


Figure 3 In the network on the left side of the figure there are two OMIMs that share two genes. The outcome inferred network links both diseases and the score is 2.

3. Semantic Similarity Relationships

Two ontologies are used to calculate semantic similarity between input elements. Gene Ontology is used to include functional relationships among genes, and the Human Phenotype Ontology allow us include phenotypic similarity among phenotypic profiles such as genes, disease or orphan diseases.

Apart from the ontology, the associations between the genes and diseases are required to calculate the semantic similarity between them. The associations between GO terms and genes are included in the annotation files provided by www.geneontology.org.

Additionally, we have used the file:

http://www.geneontology.org/gp2protein/gp2protein.human.gz

which allow associate Entrez (Geneid) genes to GO terms instead of proteins.

We use different annotations files with the Human Phenotype Ontology:

OMIM: the www.human-phenotype-ontology.org site provides the annotation file that contains relationships among diseases (OMIM and Orphan) and HPO terms. We use these annotations separately to generate two unipartite networks.

Genes Entrez: The genes_to_phenotypes.txt file has been used to get the relationships among genes and HPO terms (www.human-phenotype-ontology.org).

Input

Genes

Different gene identifiers have been included in our database. The application uses the Entrez identifier to work with genes, however the user can reference genes using other ID, such as:

Ensembl: The Ensembl names have the syntax ENSG###. The Ensembl project produces genome databases for vertebrates and other eukaryotic species. More information in http://www.ensembl.org

Example: ENSG00000169249 (ZRSR2)

HGNC or GeneSymbol: "*The HUGO Gene Nomenclature Committee (HGNC) has assigned unique gene symbols and names to over 33,000 human loci, of which around 19,000 are protein coding. http://www.genenames.org/ is a curated online repository of HGNC-approved gene nomenclature and associated resources including links to genomic, proteomic and phenotypic information, as well as dedicated gene family pages.*" *http://www.genenames.org/*

Example: 23019 (ZRSR2)

HPRD: "*The HPRD (Human Protein Reference Database) represents a centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome. All the information in HPRD has been manually extracted from the literature by expert biologists who read, interpret and analyse the published data.*" http://www.hprd.org/

Example: 02068 (ZRSR2)

MIM: "*Online Mendelian Inheritance in Man (OMIM®) is a continuously updated catalogue of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression.*" http://omim.org/

Example: 300028 (ZRSR2)

Orphanum: The Orphanum code is assigned *Thegeneslistedin* for Orp *Orphanet are those which are thought to be implicated in the pathophysiology of rare diseases. The information is extracted from the scientific literature and cross-validated*. http://www".orpha.net/

Example: 165921 (SCP2)

OMIM

All OMIM entries have been stored in our database:

"Online Mendelian Inheritance in Man (OMIM®) is a continuously updated catalogue of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression. OMIM is a continuation of Dr. Victor A. McKusick's Mendelian Inheritance in Man, which was published through 12 editions, the last in 1998. OMIM is currently biocurated at the McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University School of Medicine." http://omim.org/

In the web site http://omim.org/ is possible to search OMIM genes, phenotypes or disorders.

Orphan Disease

A disease is considered a rare disease or orphan disease when it affects one in every 2000 people. We have included in our databases the entries referred to orphan diseases provided by Orphanet. You can find more information in **http://www.orpha.net**

The list of orphan diseases has been extracted from www.orphadata.org that provide dataset related to rare diseases.

Phenotypes

The Human Phenotype Ontology provides the phenotypes included in the knowledge base. A query of phenotypes consists in a list of HPO terms.

Example: HP:0004904 (Insulin-dependent maturity-onset diabetes of the young)

More information in:
http://www.human-phenotype-ontology.org/contao/index.php

Output Networks

This process consists in getting all the relationships included in the PhenUMA database:

1. **Initial Network** includes the relationships between the input set and the rest of genes included in the database. The type of relationship considered is the requested by the user. In this case the edges represent phenotypic similarity among genes.

2. **Network enrichment:** Apart from phenotypic relationships, additional information is included to the network. Phyisical interactions, functional relationships (GO), inferred relationships from OMIM and Orphanet and metabolic relationships between the nodes of the initial network.

3. **Final Network** includes the requested relationships included in the initial network and the additional relationships.

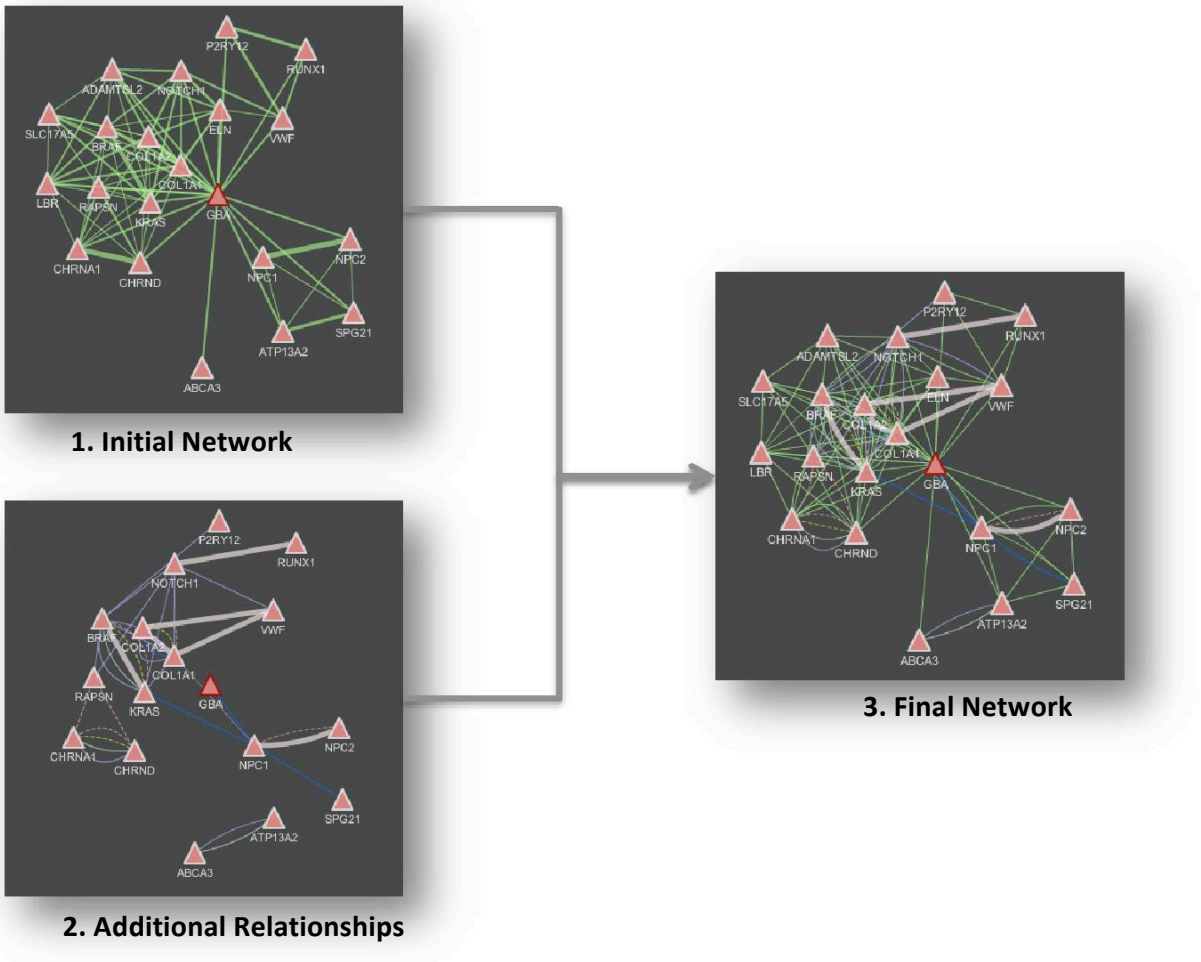


Figure 4 Network Building

.

Elements Related	Additional Relationships
Gene-Gene	Protein-Protein Interaction
	Metabolic
	Functional Similarity (Biological Process)
	Functional Similarity (Molecular Function)
	Functional Similarity (Cellular Component)
	Phenotypic Similarity
	Inferred by OMIM
	Inferred by Orphan Diseases
OMIM-OMIM	Phenotypic Similarity
	Inferred by genes
Orphan Disease – Orphan Disease	Phenotypic Similarity
	Inferred by genes
Gene – OMIM	Phenotypic Similarity
	OMIM relationships
Gene – Orphan Disease	Phenotypic Similarity
	Orphadata relationships

Table 1 This table shows the relationships included in the generated networks according to the type of elements related.

N.B. A threshold filters the semantic similarity relations that are included in each network. Each ontology and subontology has a threshold so that only include the relationships that exceed these values. In the next section specifies the cut-off selected for each type of relatioships.

Phenotyping

The phenotypes queries returns a network where the input set of phenotypes (HPO ids) is considered as a node. The resulting network shows the phenotypic relationships between the query and genes, orphan diseases or OMIM entries. The process is similar to the pervious networks. The Figure 5 shows an example of a phenotype query concretely HP:0000501 (Glaucoma):

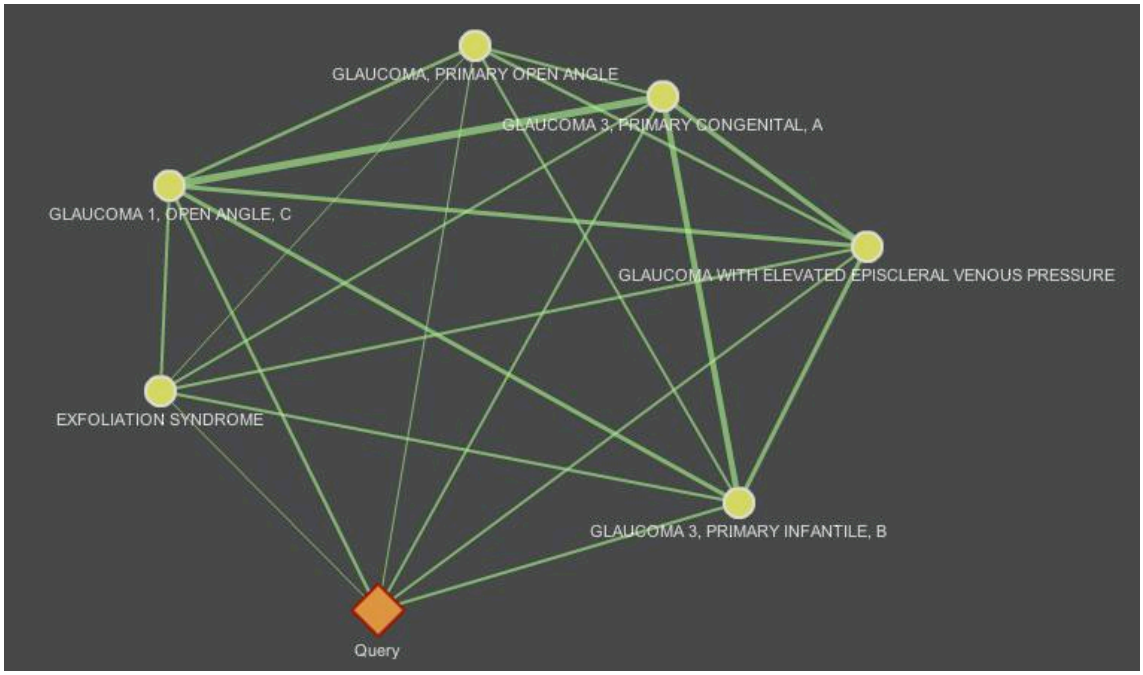


Figure 5 Resulting network of a phenotypes query.

Overview of the web interface of PhenUMA

Main Page

The main page of PhenUMA has the labeled parts in the following figure:

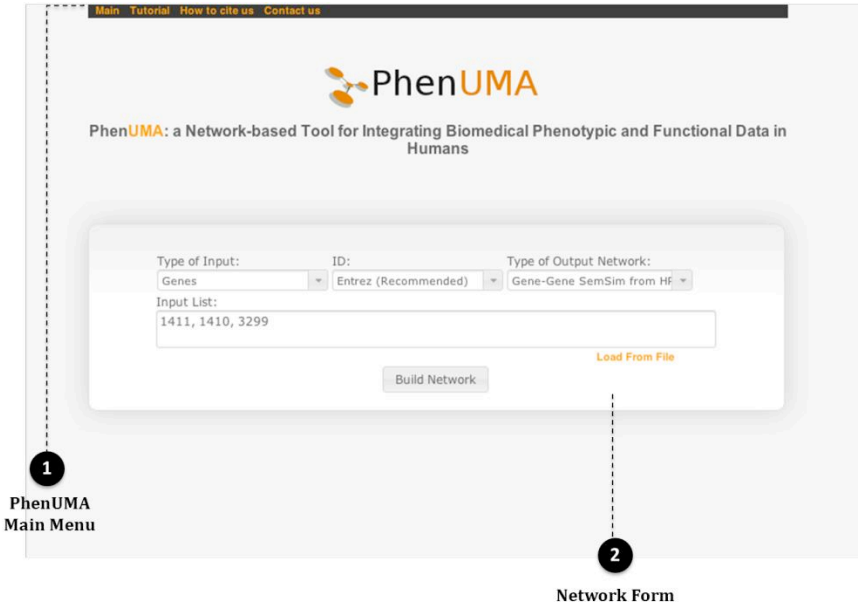


Figure 6 PhenUMA main page.

1. **PhenUMA Main Menu:** The main menu includes links to tool information: Tutorial, How to cite us and Contact us.

2. **Network Form** includes the set of fields that are requested to build a network. The allowed values of each field and the meaning will be detailed in following section of this tutorial.

Network Form

In the figure X is shown each part of the network form:

Firstly is necessary to configure the input data options and select the network of interest. In the Figure 2 we can see the following fields:

1. **Type of Input:** The allowed inputs to make a query are genes, OMIM genes/disease, Orphan diseases or Phenotypes. In this field the type of element used must be selected.

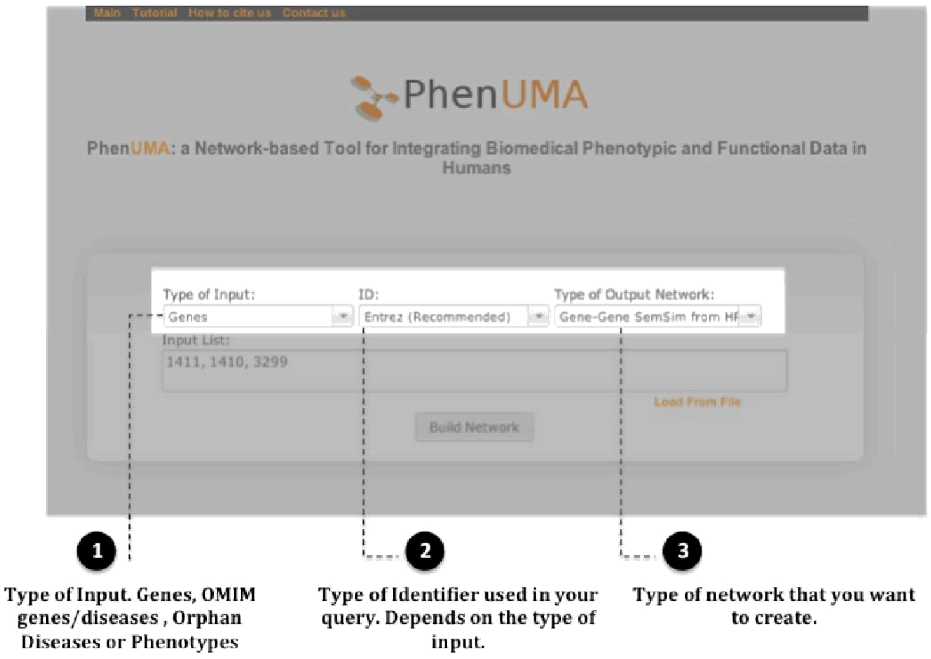


Figure 7 Network form fields

2. **ID:** Identifier used to represent the elements of the input list. The following table shows the identifiers available for each type of input:

Input Type	ID	URL
Genes	Entrez	http://www.ncbi.nlm.nih.gov/gene
	Ensembl	http://www.ensembl.org/index.html
	GeneSymbol	http://www.genenames.org/
	HGNC	http://www.genenames.org/
	MIM	http://omim.org/
	Orphanum	http://www.orpha.net/consor/cgi-bin/index.php
OMIM Genes/Diseases	OMIM	http://omim.org/
Orphan Diseases	Orphanum	http://www.orpha.net/consor/cgi-bin/index.php
Phenotypes	HPO id	http://compbio.charite.de/phenexplorer/

Table 2 Identifiers used for each type of input

3. **Type of output network**: In the third field the type of network of interest must be selected. The following graphs show the type of output networks available for each type of input.

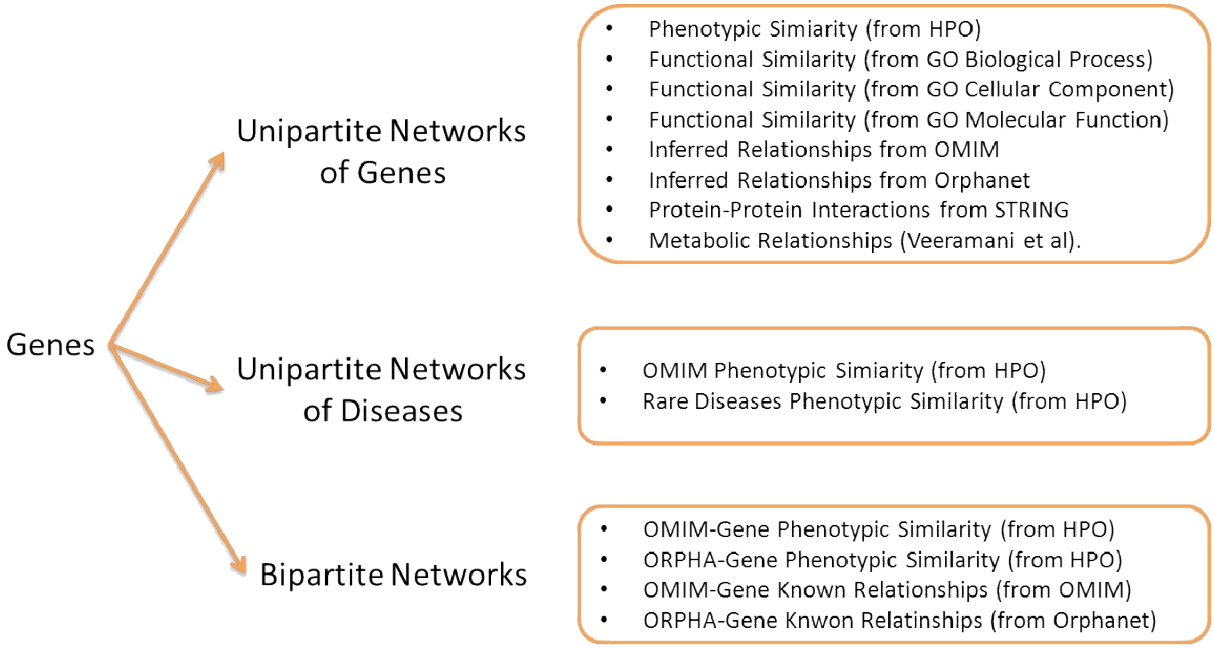


Figure 8 Output network for genes queries

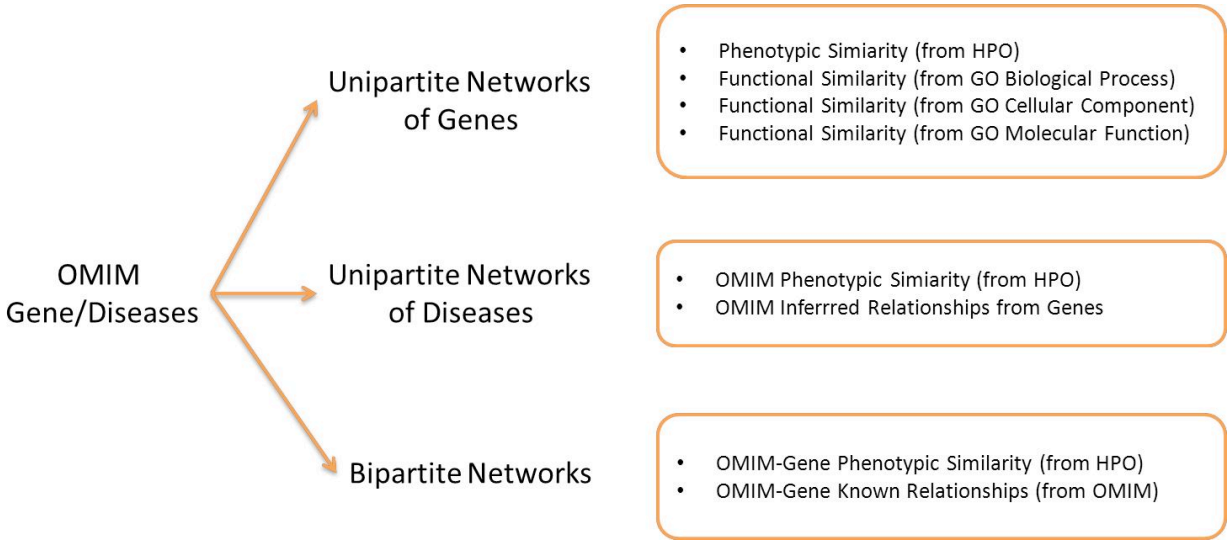


Figure 9 Output network for OMIM networks

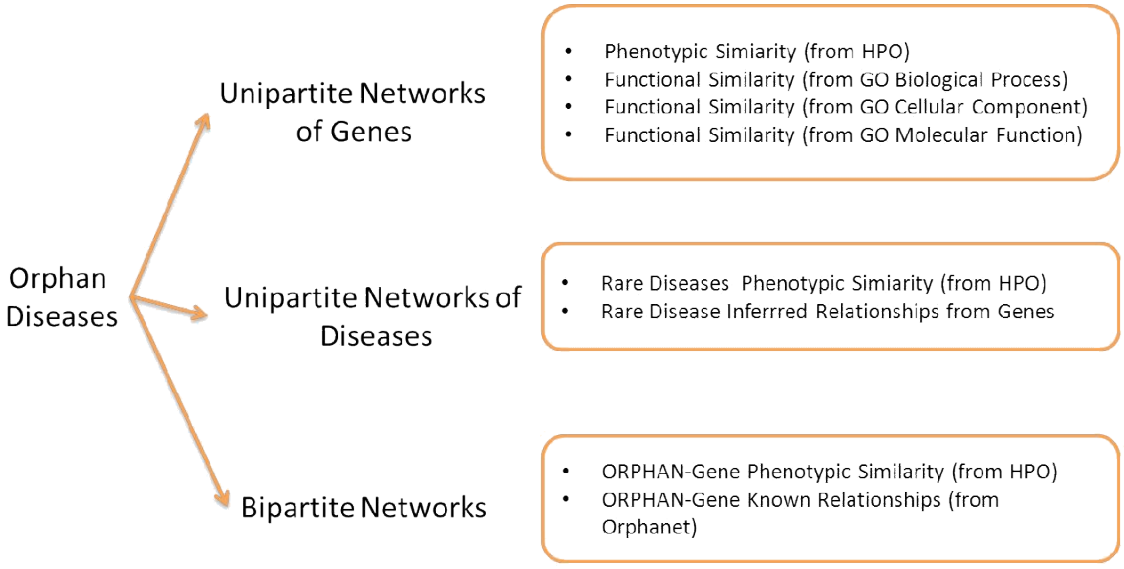


Figure 10 Output networks for Orphan Diseases queries

If the input list is a set of ***Phenotypes*** the Types of Output Network are:

OMIM (Gene/Diseases)
Orphan Diseases
Genes

In this case the output network is composed of the relationships between the input query (phenotypes) and the selected object type.

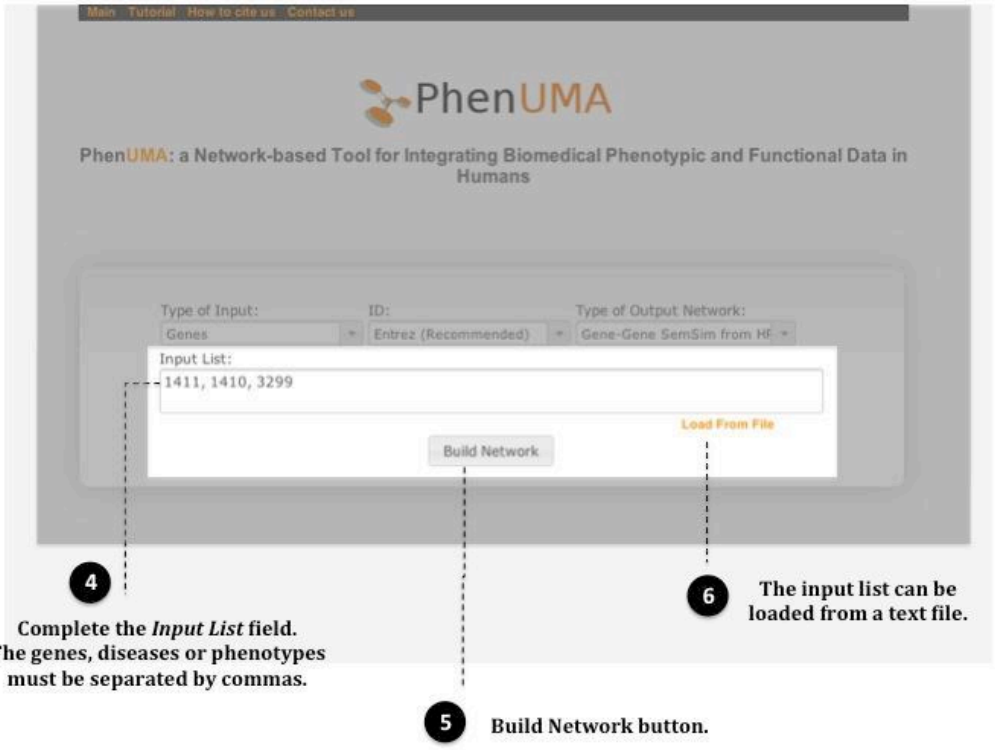


Figure 11 Network form fields: Input list

Once the type of query has been configured the **Input List** field must be completed as the next image shown:

4. **Input List:** The list of object of your query must be typed to the Input List field. Notice that the identifiers used must match the selected option in the previous step (ID field) and be separated by commas. As you can see in the snapshot the input set is composed by three genes (1411,1410, 3299) identified by the Entrez code.

5. When the **Build Network** is pressed the network will start to be created. This process could take some minutes. That depends of the size of the network requested. A progress bar will be shown while the network is built.

6. **Load From File:** The input list can be filled from a text file. Pr label and in the dialog form pres *Choose* to select the file and *Upload* to include the content of your file in the Input List field.

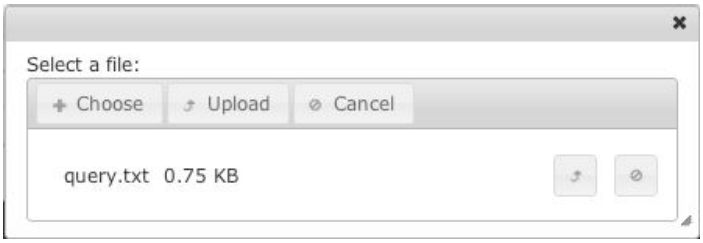


Figure 12 Load from file dialog

When the network is built without problems, PhenUMA show the display with the requested network. If the display is not shown may be due to two possibilities:

The network is empty: this message is shown when there is not information enough to build a network.

The network is too big: this message appears when the resulting network is too large for the visualization tool. In this case, the link is available, so you can download the network in a text file format and use other software as Cytoscape.

Display Page

When a network is requested in the *Main Page,* PhenUMA gets the information from the database and built the network, then the *Display Page* is shown and the network is displayed as shown the next figure:

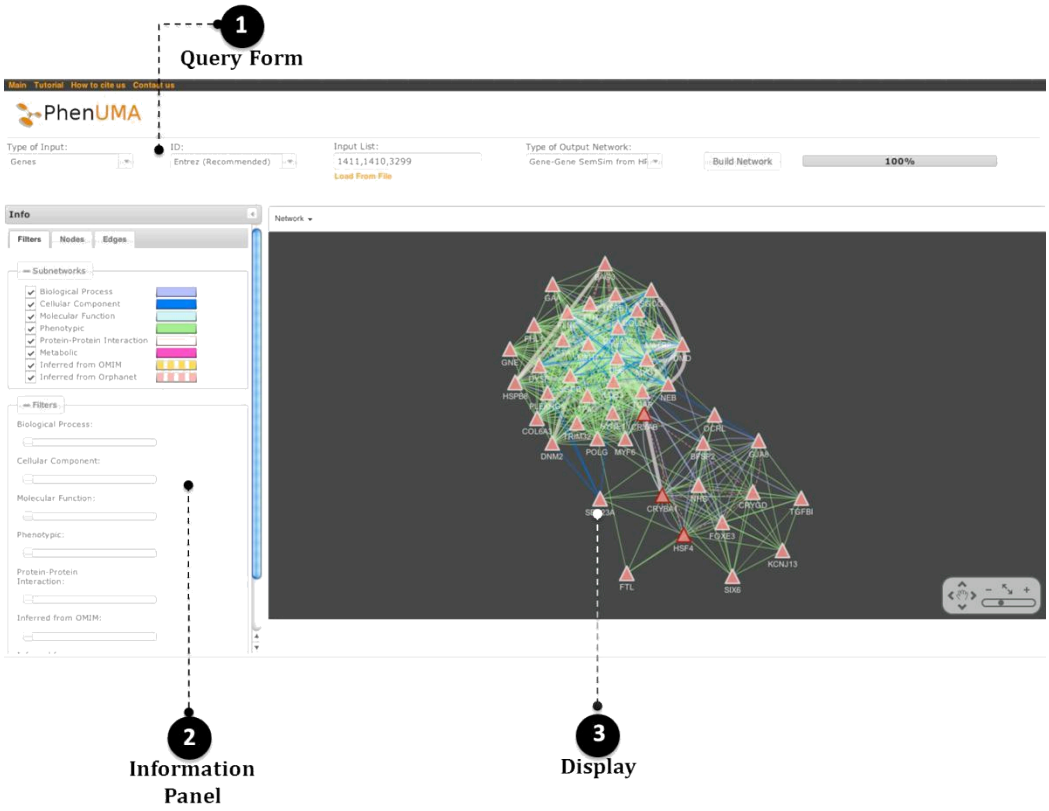


Figure 13 Display page

As it observed, the *Visualizer Page* is formed by the following sections:

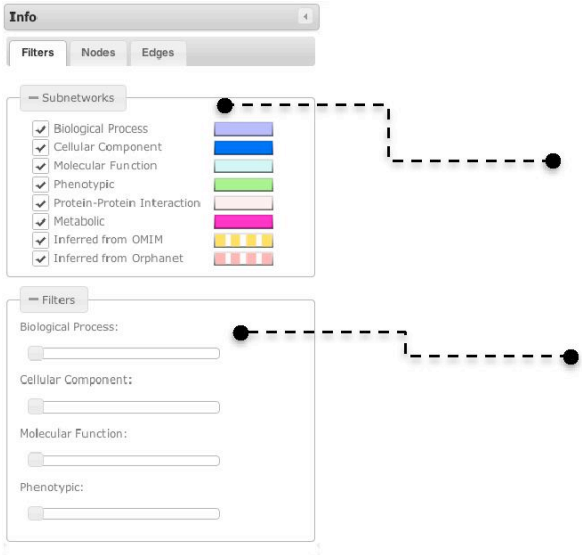
1. Query form

The visualizer page includes the form used to query a network with the same fields of the Main Page, to execute other queries.

2. Information Panel

The information panel has three tabs: **subnetworks**, **nodes** and **edges**.

A. Filters



Subnetworks: This panel shows the color code used to display the relationships of a network. You can hide/show the relationships using the checkbox placed next to each relationship type name.

Filters: The relationships included in the output network can be filtered to exclude the relationships less significatives.

Figure 14 Information Panel

B. Nodes

This tab has two tables. The upper table shows the nodes selected in the display. The bottom table shows the phenotypes (HPO terms) of the node selected in the upper table and includes the HPO Id, Name and IC of each phenotype.

The IC value is the informativeness of the phenotype and indicates the specifity. Phenotypes with value of IC close to 1 are very specific while the closer to zero more genera.

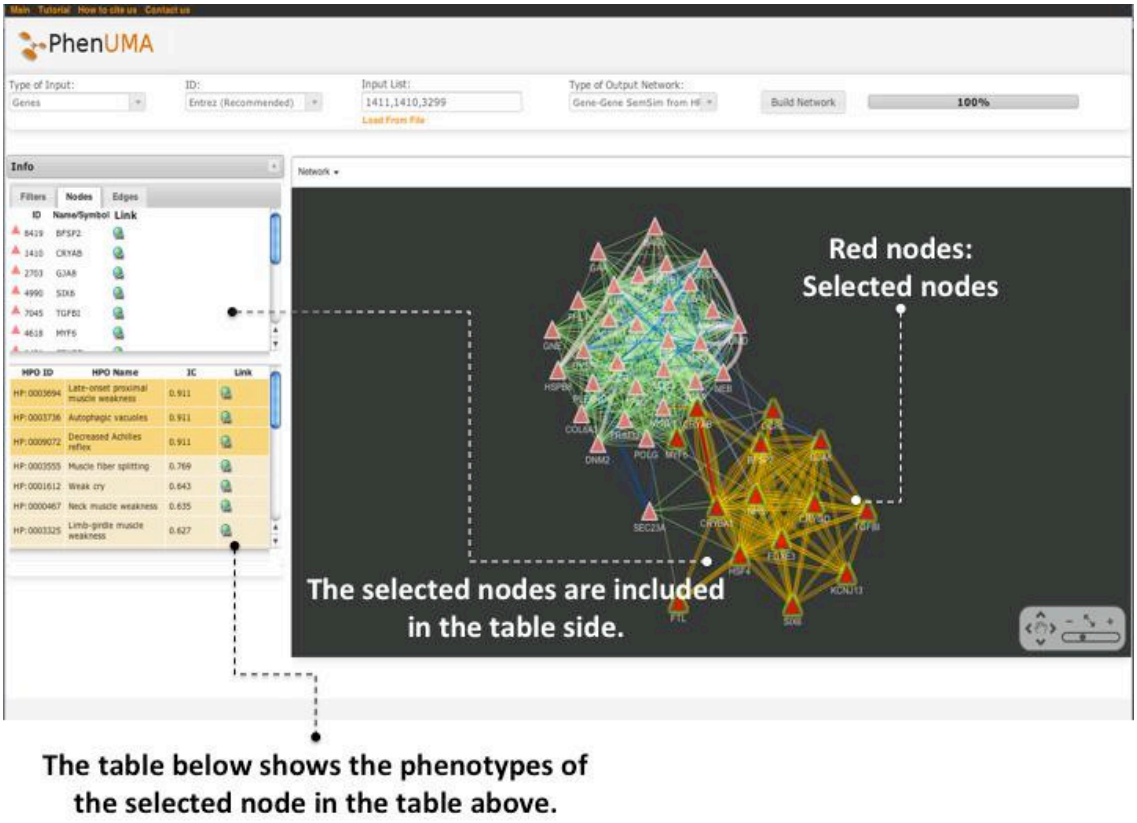


Figure 15 Information Panel : Nodes Tab

C. Edges

The edges tab has two tables. The upper table shows the edges selected in the display and for each relationship is included the nodes connected, the score and the type. When a row of the *relationship table* is selected in the bottom table shows additional information of the relationship, and this information depends of the type of this relationship.

Phenotypic relationships: the common phenotypes of the genes or diseases related.

Functional relationships: the common go terms of the genes related.

Inferred relationships: in case of a gene-gene relationships the common diseases of the genes and in case of a omim-omim or orphan-orphan relationships the common genes of the diseases.

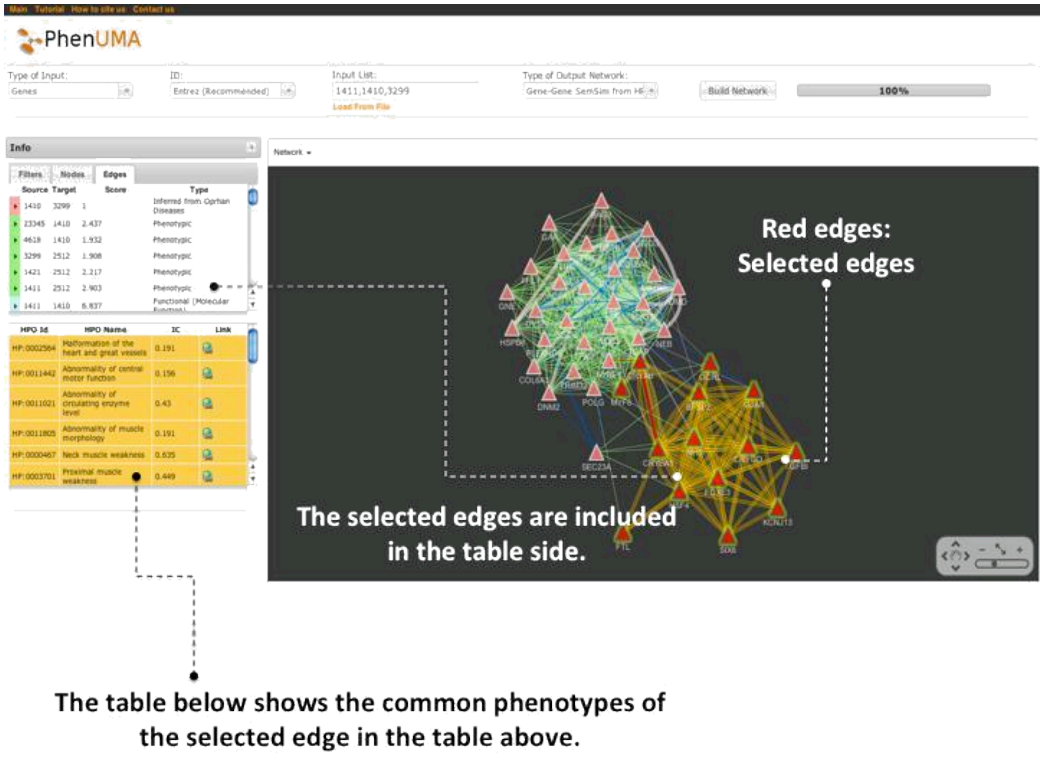


Figure 16 Information Panel : Edge Tab

3. Network Display

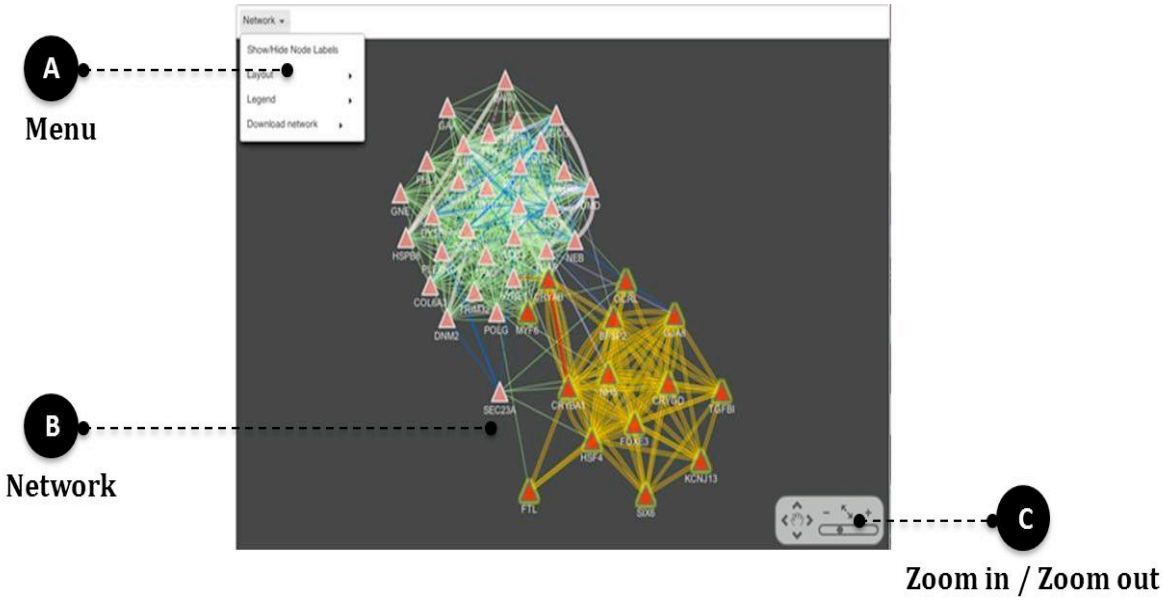


Figure 17 Network Display

A. Menu: The options of the network menu are:

Show/Hide nodes labels.

Apply Layout: With this option is possible to organize the nodes in different configurations. The next figure shows the supported possibilities by PhenUMA. The layout can be applied from the visualizer menu.

Network → Layout

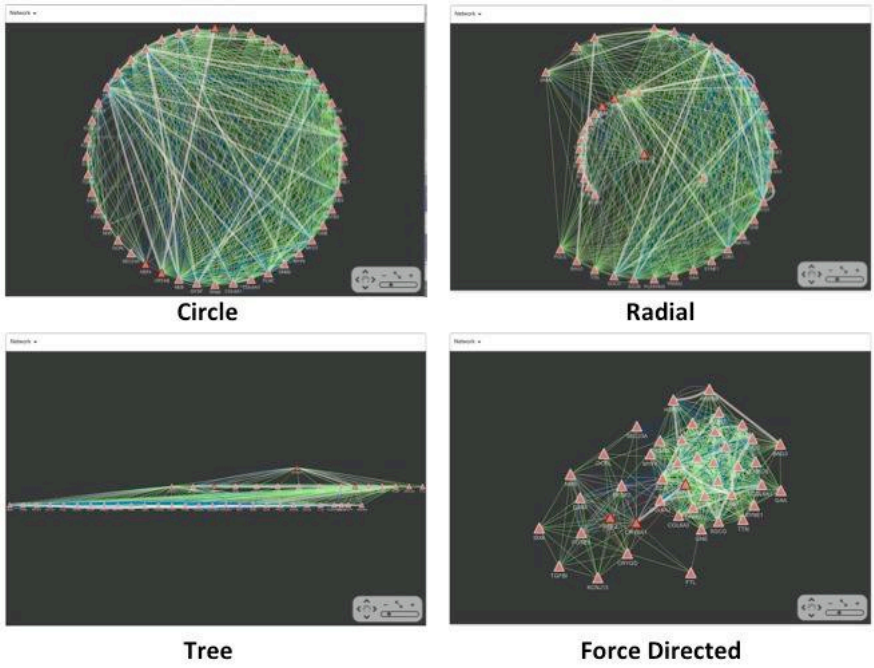


Figure 18 Layouts

Legend: This option shows an image with the meaning of each shape used in the networks.

Download the network in a text file: The outcome network can be downloaded from the visualizer menu:

Network → Download Network → Text File.

Each relationships is represented in the file as a row with four columns: Node source
Node target Score
Relationship type**.**

The next image shows an example. A network and the corresponding text file:

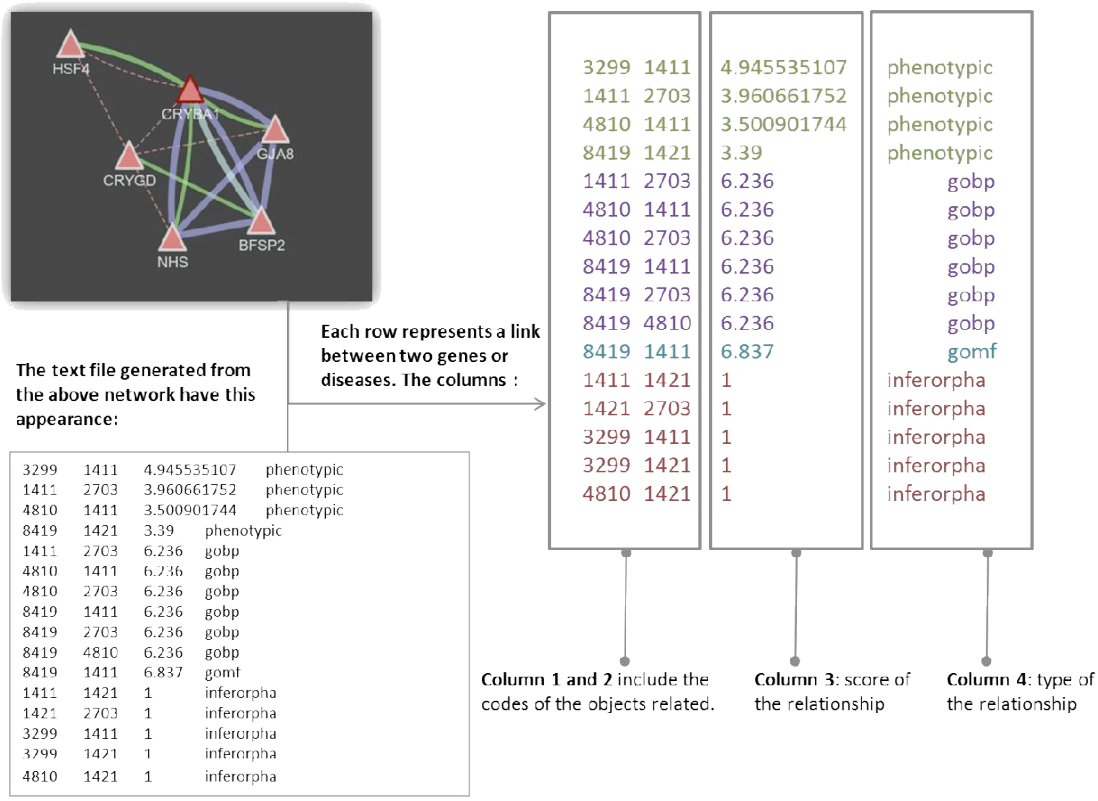


Figure 19 Text File Example

A color code is used in PhenUMA to represent each type of relationships. In this example the green links represents phenotypic relationships, the purple and blue relationships represent functional relationships from Gene Ontology (biological process and molecular function respectively) and the red dashed bonds indicate that the related genes are involved in the same orphan disease. The rest of relationships and their representation will be detailed in the following section of this tutorial.

The column 4 is used to indicate the type of relationship. The label used in the text file to represent each type of relationship is:

Phenotypic from HPO : **phenotypic**
Functional from GO (Biological Process) : **gobp**
Functional from GO (Molecular Function) : **gomf**
Functional from GO (Cellular Component): **gocc**
Protein-Protein Interaction : **ppi**
Metabolic : **metabolic**
Gene-Gene inferred from OMIM : **inferomim**
Gene-Gene inferred from Orphanet : **inferorpha**
OMIM-OMIM inferred from Gene : **infergene**
Orpha-Orpha inferred from Gene : **infergene**
OMIM-Gene from OMIM : **omim**
Orphan Disease – Gene from Orphadata: **orphadata**

Each text file includes a header where is detailed the meaning of each label used, for example, the following header is included in any unipartite network of genes:

```
# Relationships:
# - gobp : Biological Process (Gene Ontology)
# - gocc : Celullar Component (Gene Ontology)
# - gomf : Molecular Function (Gene Ontology)
# - phenotypic : Phenotypic Similarity (Human Phenotype Ontology)
# - ppi : Protein-Protein Interaction (STRING)
# - metabolic : Metabolic Flux Correlation
# - inferomim : Inferred Relationship by OMIM. These genes are related with the same set of OMIM diseases.
# - inferorpha : Inferred Relationship by Orphan Diseases (Orphadata). These genes are related
```

B. Network: PhenUMA uses the CytoscapeWeb tool to display the networks. This visualizer allows:

Select Nodes and Edges: the nodes and edges of the displayed network can be selected clicking on them with the mouse. The selected nodes/edges will be highlighted and the information related with them will be shown in the side panel. The displayed data in this panel will be detailed in the next point.

Move nodes or groups of nodes: When a node o group of nodes is selected you can move them.

Select the neighbors of a node by double clicking on a node.

The networks are built using different shapes and colors for every type of object: Genes, OMIM, Orphan Diseases and Phenotypes Queries. The inputs elements are highlighted to make easier distinguish them from the rest of nodes. We use different colors to represent the relationships too. The following schema shows the color/shape code:

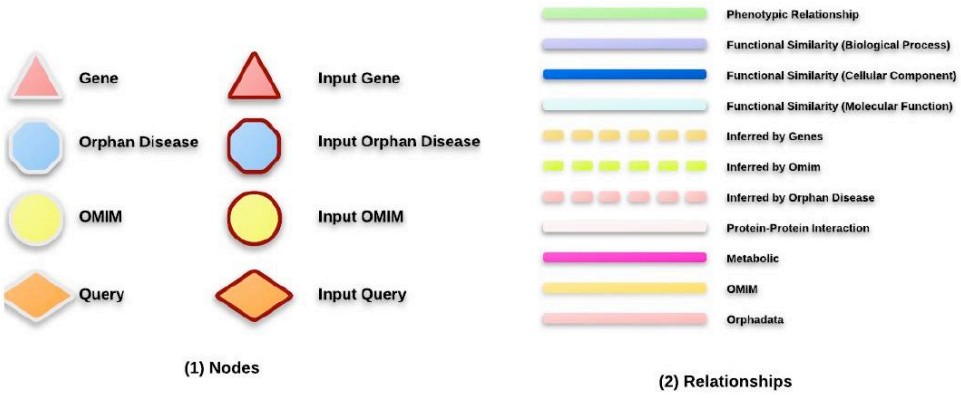


Figure 20 Color/Shape code

C. Zoom in / Zoom out.