Phenolyzer Manual

Phenolyzer is a tool to help prioritize disease genes based on any disease/phenotype terms as input. The simplest input into Phenolyzer is just a description, like 'Alzheimer'. And the output is a list with prioritized genes, scores and all the details.

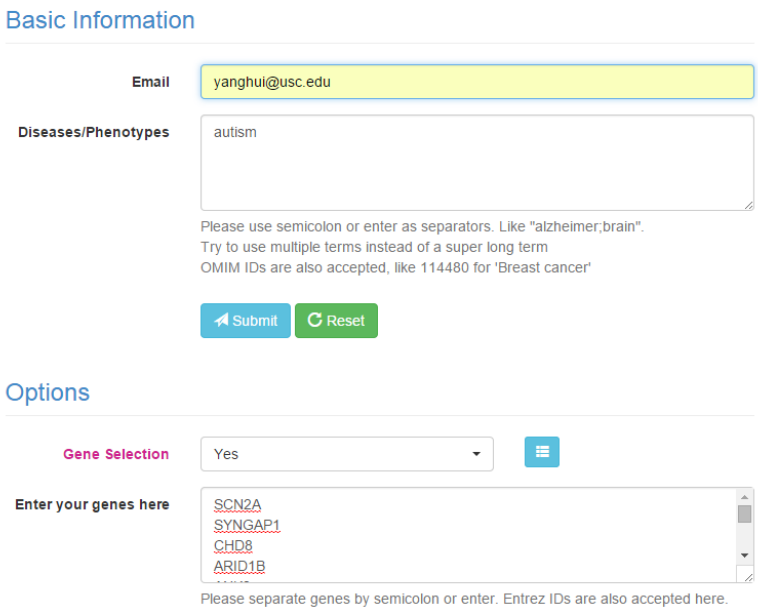
Quick Start: End-to-End example

To give a quick example of how to use Phenolyzer and how Phenolyzer works, one quick example is given here.

Suppose you are a scientist studying genetics about autism, and you conducted an exome sequencing on a patient and got 1000 variants after all the data processing. Now you face a reality that 1000 variants are too many for you to figure out which variant is worth for the following study.

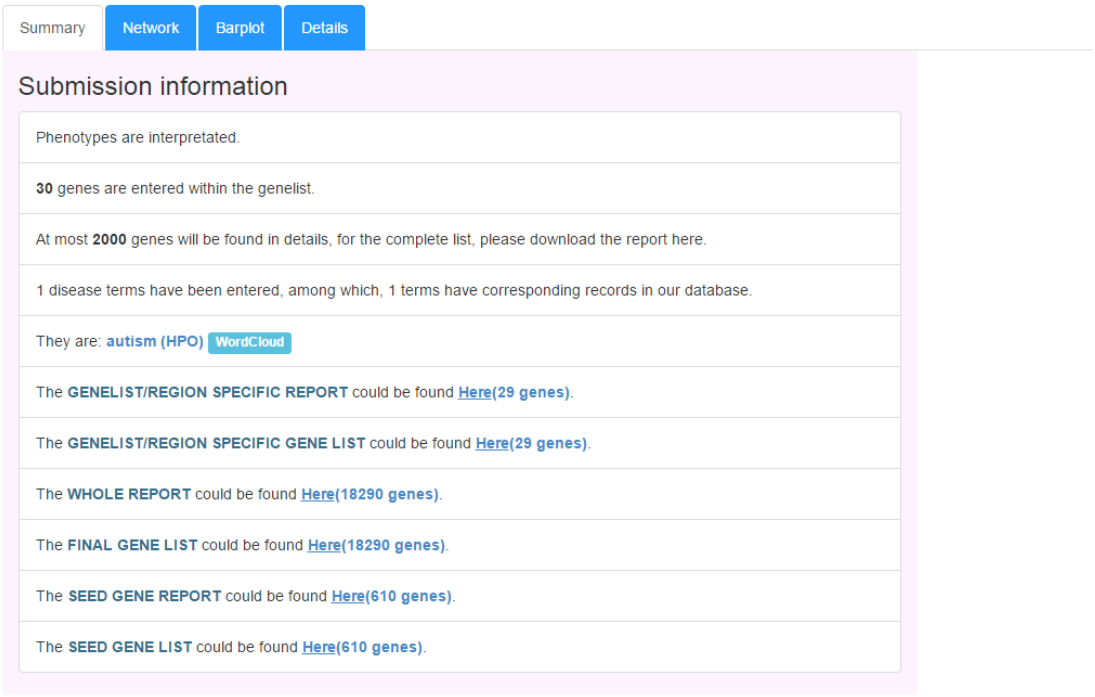
What can you do? You can search all the literature and databases but probably they won't give you a direct and intuitive order of the importance of these genes for autism.

Then Phenolyzer comes! Just type in 'autism' in at http://phenolyzer.usc.edu/, turn on the Gene Selection option, and paste your gene list here, and submit. Then… it's done!



After 1 minute, we got our result page:

We have our network, our result text files for download, our bar plot and our detail section:



To understand how each score is generated for each gene, let's take a simple journey here.

First, 'autism' is tried to match all the disease and phenotype names in our database. From the disease details by clicking 'autism' link in Summary section, we got a text file. From the text file, we can find records like

```
angelman syndrome autism spectrum disorder  GENE_DISEASE
Rett Syndrome, Zappella Variant CTD_DISEASE
secretory diarrhea autism   GENE_DISEASE
hypoadrenalism autism x linked 5    GENE_DISEASE
attention deficit hyperactivity disorder autism GENE_DISEASE
autism spectrum disorder specific language impairment 5 GENE_DISEASE
MENTAL RETARDATION WITH LANGUAGE IMPAIRMENT AND AUTISTIC FEATURES   CTD_DIS
ASPGX1;ASPERGER SYNDROME, X-LINKED, SUSCEPTIBILITY TO, 1    CTD_DISEASE
cortical dysplasia focal epilepsy syndrome  0.33
15q14 microdeletion syndrome    0.075
```

'autism spectrum disorder specific language impairment' is contains 'autism', thus it is one kind of autism and this disease name is saved. 'ASPGX1;ASPERGER SYNDROME, X-LINKED, SUSCEPTIBILITY TO, 1' has no autism in it, but it is an offspring disease of 'Autism spectrum disorder' in CTD Disease Vocabulary thus it is also saved. '15q14 microdeletion syndrome' does not contain 'autism' either, but it is related with autism phenotype, thus also saved with a score. Together, these are all the diseases related with autism and will be queried in our disease-gene database.

Also another file contains all the HPO terms like this:

```
HP:0000717
HP:0000723
HP:0000728
HP:0000729
HP:0000735
HP:0000753
HP:0000758
HP:0000817
HP:0002332
HP:0008763
```

This is just the HPO terms whose names contain ‘autism’ as a word.

After the database lookup for all the disease names we got at previous step, we got our seed genes and the details like this:

```
Tuple number in the gene_disease databse for all the terms: 1050
MECP2   ID:4204 -       1
unknown (GENE_CARDS)    asperger syndrome       autism  0.0377977947477198
OMIM:300496 (OMIM)      autism  autism  0.151191178990879
OMIM:300496 (CLINVAR)   autism  autism  0.00944944868692994
umls:C1845336 (DISGENET)        autism  autism  0.0453573536972638
unknown (GENE_CARDS)    autism  autism  0.0377977947477198
umls:C0004352 (DISGENET)        autistic disorder       autism  0.0509082951980014
umls:C0008074 (DISGENET)        child development disorders pervasive   autism  0.0453573536972638
unknown (GENE_CARDS)    epileptic encephalopathy early infantile 2      autism  0.0377977947477198
umls:C1846058 (DISGENET)        lubs x linked mental retardation syndrome       autism  0.0150721986860
unknown (GENE_CARDS)    lubs x linked mental retardation syndrome       autism  0.0124732722667475
unknown (GENE_CARDS)    mental retardation x linked     autism  0.0377977947477198
```

We can see that MECP2 has the highest normalized score, as it has a record in OMIM with ‘autism’, corresponding to the term ‘autism’, with a score ‘0.15119…’. It also has a record in CLINVAR to be associated with autism, also in GENECARDS, DISGENET and so on. Then this normalized score is treated as one feature.

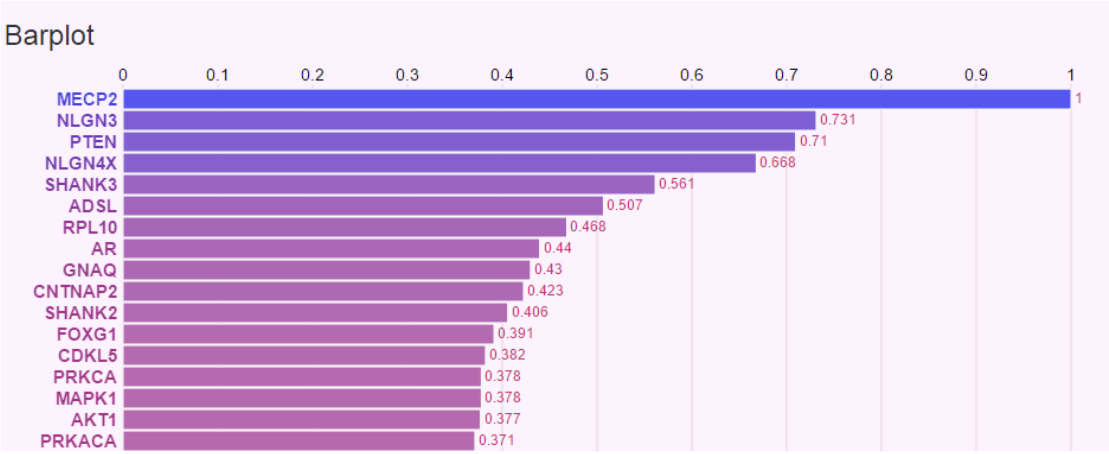
Each gene in the whole genome which has an interaction with these seed genes will also get an additional score, corresponding to four kinds of interactions: protein interaction, in the same Biosystem, in the same Gene Family and in transcription interaction. All these four scores are treated as four other features.

So each gene has 5 features now! Our machine learning model is trained with all these features, to classify the gene as disease gene or unrelated genes, and the trained parameters are used to calculate a score. Finally these calculated raw scores are normalized again and we got a final list and detailed report, like this:

1 PTEN	SeedGene	Raw Score:6.65
PTEN		
umls:C0004352 (DISGENET)	autistic disorder	autism (0.6123)
umls:C0008074 (DISGENET)	child development disorders pervasive	autism (0.4324)
OMIM:158350 (CLINVAR)	cowden syndrome	autism (0.01351)
ORPHANET:201 (ORPHANET)	cowden syndrome	autism (0.09266)
OMIM:605309 (OMIM)	macrocephaly autism syndrome	autism (1.441)
OMIM:605309 (CLINVAR)	macrocephaly autism syndrome	autism (0.09008)
ORPHANET:210548 (ORPHANET)	macrocephaly autism syndrome	autism (1.235)
umls:C1854416 (DISGENET)	macrocephaly autism syndrome	autism (0.8648)
unknown (GENE_CARDS)	macrocephaly autism syndrome	autism (0.3603)
PUBMED:12620407 (HPRD)	in vivo;in vitro	With TP53 (0.02431)
PUBMED:15205473 (HPRD)	in vivo;in vitro	With AR (0.03325)
PUBMED:12077256 10866658 12297295 (HPRD)	in vivo;in vitro	With AKT1 (0.008102)
BIOSYSTEM:530764 (BIOSYSTEM)	In the same (Disease)	With RPL10 (0.03429)
BIOSYSTEM:477114 (BIOSYSTEM)	In the same (Signal Transduction)	With FSHB (0.003318)
BIOSYSTEM:477135 160976 219801 (BIOSYSTEM)	In the same (Metabolism; Metabolism of lipids and lipoproteins; Integrated Breast Cancer Pathway)	With HMGCR (0.003318)
BIOSYSTEM:106387 833825 106386 (BIOSYSTEM)	In the same (Innate Immune System; Fc epsilon receptor (FCERI) signaling; Immune System)	With RASGRP4 (7.901e-05)

Where PTEN has a record in disease-gene mapping step thus is a seed gene, and its Raw Score is 6.65. It is scored as 6.65 because it has the disease mappings in DISGENET, ORPHANET, OMIM, CLINVAR and all these databases. At the same time, it interacts with TP53, AR, AKT1 and other seed genes. All the scores corresponding to the details should sum up to the Raw Score.

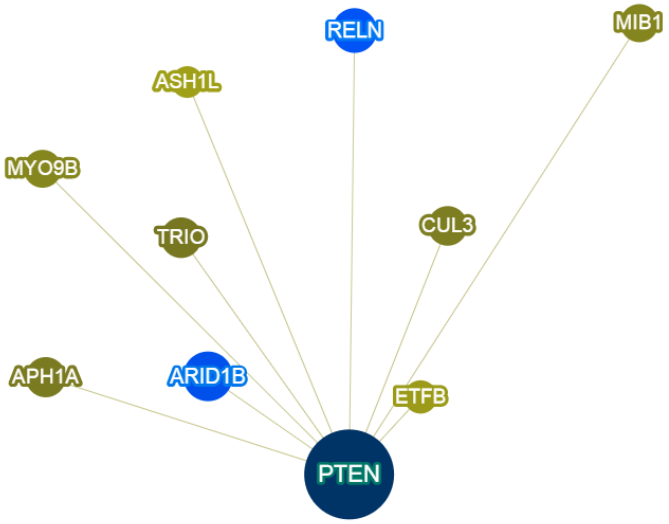
For the normalized scores, it is like this:



To visualize this process, we can look at the Network. First the term ‘AUTISM’ is mapped to different disease names, like ‘autism’, ‘coffin siris syndrome’, and ‘mental retardation autosomal dominant 7’.

Then, each disease name is mapped into some seed genes, like coffin siris syndrome is mapped into ARID1B.

Finally, the seed genes are grown and all the associated genes are returned, like PTEN is in the same Biosystem with RELN, CUL3, ETFB, and so on:



Prepare Input

Phenolyzer takes any disease/phenotype terms as input, including:
1) full disease names, like ‘amyotrophic_lateral_sclerosis’, ‘autism spectrum disorder’, ‘Spinocerebellar Ataxia 38’, and so on. The autocomplete in Phenolyzer website can be used to check the availability of the long names. If there is no record in the autocomplete, then it is unlikely that there will be a record and the short terms should be considered.
2) OMIM ID: like 114480 for 'Breast cancer', OMIM IDs can be retrieved in OMIM website at http://www.omim.org/
3) Phenotype terms: like ‘fatigue’, ‘short stature’, ‘cough’ and so on.
4) HPO terms: HPO terms corresponding to a disease can be queried with Phenomizer, at http://compbio.charite.de/phenomizer/
5) general disease terms: like ‘cancer’, ‘coronary disease’, ‘diabetes’ and so on.
6) specific full disease names: like ‘Spermatogenic failure nonobstructive Y-linked’.

The strategy to prepare your input terms should depend on the need. If the need is to get as many genes as possible then more general and shorter terms should be used. Otherwise exact and full disease names should be considered.
If the diagnosis is uncertain, then phenotype terms should be used. Both HPO terms and general descriptions can be considered.

Warning: There should be no non-word characters within your terms except commas. Other delimiters might make Phenolyzer treat your input as different terms.

If multiple disease or phenotype are simultaneously considered, then multiple terms should be input, delimited by ‘;’ or Enter. One tip here is that if the disease name is too long, then it can be artificially separated into multiple terms, to be more generative. Phenolyzer itself won’t help you tokenize your input into multiple terms.

Other types of input include 1) Gene list, 2) genomic region file in BED format. The Gene list will make Phenolyzer generate another file, which only prioritize the genes within the given list. The genomic region confines Phenolyzer to prioritize the genes within the given regions. Also the genomic region will be treated as Copy Number Variations (CNV) and these CNVs will also be prioritized.

Understand Output

There are three types of text output generated by Phenolyzer: 1) gene list, 2) detail file and 3) disease/phenotype interpretation file.

For the gene list, it is in the format as below:

Rank	Gene	ID	Score	Status
1	SLC7A7	9056	1	SeedGene
2	PRF1	5551	0.8547	SeedGene
3	UNC13D	201294	0.7097	SeedGene
4	STX11	8676	0.6178	SeedGene
5	STXBP2	6813	0.5757	SeedGene
6	XIAP	331	0.4723	SeedGene
7	RAB27A	5873	0.4282	SeedGene
8	MYH9	4627	0.3995	SeedGene
9	GATA1	2623	0.3899	SeedGene
10	TNF	7124	0.3493	SeedGene
11	ITGB3	3690	0.343	SeedGene
12	CD27	939	0.3233	SeedGene
13	NLRP3	114548	0.3208	SeedGene
14	ITGA2B	3674	0.3071	SeedGene
15	IL6	3569	0.3005	SeedGene
16	IFNG	3458	0.2995	SeedGene
17	AKT1	207	0.2955	Predicted
18	IL1B	3553	0.2945	SeedGene
19	GATA2	2624	0.2917	SeedGene
20	ICAM1	3383	0.2802	SeedGene
21	TNFRSF1A	7132	0.28	SeedGene
22	THBD	7056	0.279	SeedGene
23	MPL	4352	0.277	SeedGene

Rank: The rank of each gene, which does not consider the situation where some genes have the same score.
Gene: The Entrez gene name.
ID: The Entrez gene ID.
Score: The normalized score for each gene. The range of the scores is from 0 to 1.
Status: If the gene is a seed gene or a purely predicted gene. A seed gene means this gene has some direct relations with the input term, based on the existing databases. A purely predicted gene means the gene is found by gene-gene relations.

There is also a seed gene list which only contains seed genes, and has normalized scores from 0 to 1 for all the seed genes.

For the details, it is in the format as below:

```
Tuple number in the gene_disease database for all the terms: 1050
MECP2   ID:4204 - SeedGene  9.99127878528306    Normalized score: 1
unknown (GENE_CARDS)    asperger syndrome   autism  0.36033380837646
OMIM:300496 (OMIM)  autism  autism  1.44133523350584
OMIM:300496 (CLINVAR)   autism  autism  0.090083452094115
umls:C1845336 (DISGENET)    autism  autism  0.432400570051752
unknown (GENE_CARDS)    autism  autism  0.36033380837646
umls:C0004352 (DISGENET)    autistic disorder   autism  0.485318786693383
umls:C0008074 (DISGENET)    child development disorders pervasive   autism  0.432400570051752
unknown (GENE_CARDS)    epileptic encephalopathy early infantile 2  autism  0.36033380837646
umls:C1846058 (DISGENET)    lubs x linked mental retardation syndrome   autism  0.143686233268689
unknown (GENE_CARDS)    lubs x linked mental retardation syndrome   autism  0.118910156764232
unknown (GENE_CARDS)    mental retardation x linked autism  0.36033380837646
OMIM:312750 (CLINVAR)   rett disorder   autism  0.090083452094115
OMIM:312750 (OMIM)  rett syndrome   autism  1.44133523350584
ORPHANET:778 (ORPHANET) rett syndrome   autism  1.02952516678989
OMIM:312750 (GENE_REVIEWS)  rett syndrome   autism  0.72066761675292
umls:C0035372 (DISGENET)    rett syndrome   autism  1.29720171015526
PUBMED:19495527 19495527 (GAD)  rett syndrome   autism  0.00857937638991571
unknown (GENE_CARDS)    rett syndrome   autism  0.36033380837646
PUBMED:15696166 (HPRD)  in vivo;in vitro    With SMARCA2    0.037701066686364
PUBMED:1569166 (HPRD)   in vivo;in vitro    With SMARCB1    0.0807296027412471
PUBMED:20610535 (HTRI)  Chromatin Immunoprecipitation coupled with microarray   Regulated by AR 0.0272591330111929
BIOSYSTEM:198901 (BIOSYSTEM)    In the same (SIDS Susceptibility Pathways)  With HTR2A  0.00882729121056259
BIOSYSTEM:198901 (BIOSYSTEM)    In the same (SIDS Susceptibility Pathways)  With DLX2   0.00813069240188015
BIOSYSTEM:198901 (BIOSYSTEM)    In the same (SIDS Susceptibility Pathways)  With TF 0.00796431587864403
```

The first line of the file shows the sum of the number of disease-gene mappings corresponding to each term. From the second line, the detail section for each gene is listed, separated by a blank line.

For each section, the first line shows the gene name, ID, position (if a region file is in the input and the gene is in the region, otherwise it is a '-'), raw scores and normalized scores. Then the details are shown below.

The first part of the details are for the disease-gene mappings, which are in the 'Reference' 'Full Disease Name' 'Term ''Raw Score' format. After this is the gene-gene mapping details, which are in the 'Reference ''Mapping details' 'Related Genes ''Raw Score' format.

The way to differentiate these two types of information is by checking the third column: if it is a disease/phenotype term there, then it is a disease-gene mapping detail, otherwise it is a gene-gene mapping detail. The sum of all the scores in the detail section should be equal to the raw score of the gene.

For disease/phenotype interpretations, two files corresponding to each term are available.

The first file is the interpreted diseases, which is like below:

```
secretory diarrhea autism   GENE_DISEASE
hypoadrenalism autism x linked 5    GENE_DISEASE
attention deficit hyperactivity disorder autism GENE_DISEASE
autism spectrum disorder specific language impairment 5 GENE_DISEASE
MENTAL RETARDATION WITH LANGUAGE IMPAIRMENT AND AUTISTIC FEATURES   CTD_DISEASE
ASPGX1;ASPERGER SYNDROME, X-LINKED, SUSCEPTIBILITY TO, 1    CTD_DISEASE
cortical dysplasia focal epilepsy syndrome  0.33
15q14 microdeletion syndrome    0.075
myhrs   0.33
pthsl1 included 0.33
ichthyosis x linked 0.075
chromosome 17p11 2 duplication syndrome 0.33
periventricular nodular heterotopia 2   0.33
Chromosome 3q29 microdeletion syndrome  0.0645161290322581
```

Each full disease name corresponding to the input term is here. The synonyms are separated by semicolon. If the disease is interpreted directly by disease ontology or synonyms, the source is in the second column, like 'CTD_DISEASE' means this full name is retrieved from CTD Disease Vocabulary, 'DISEASE_ONTOLOGY' means it is from Disease Ontology database, 'GENE_DISEASE' means it matches a disease name in our pre-compiled disease-gene mapping databases.
If the disease if interpreted solely by phenotype-disease mapping, then a score is in the second column, which represents the reliability of this interpretation.

If the term has corresponding Human Phenotype Ontology (HPO) terms, a file containing HPO IDs is also available, with a single HPO ID as each line in it.

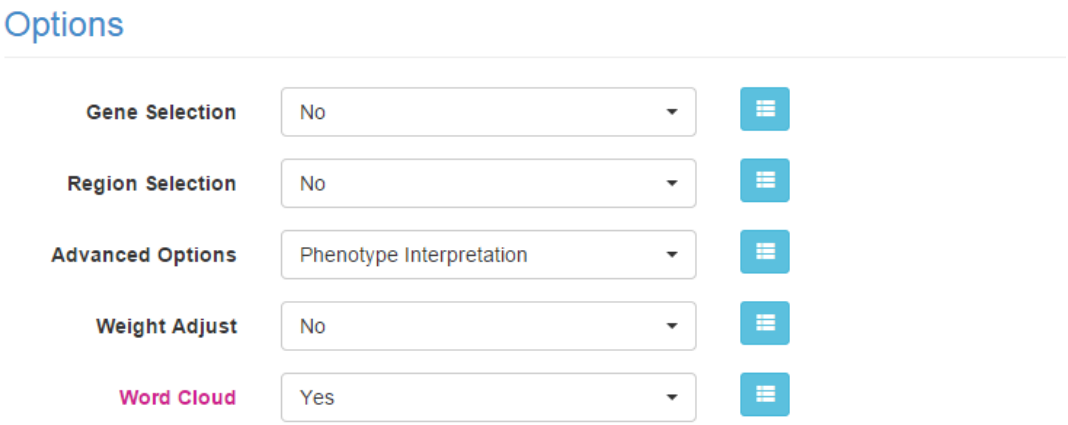
Web tool instruction

The recommended way of using Phenolyzer is through visiting its web server at http://phenolyzer.usc.edu. All of Phenolyzer's functions are realized here. The user needs to fill in their email and disease/phenotype terms.

The gene list can be entered by selecting 'Yes' for Gene Selection. The Region file can be entered by selecting 'Yes' for Region Selection.

If no Phenotype interpretation is wanted, it can be turned off by choosing 'Disease Only' in Advanced Options. If no specific disease/phenotype term is available, then 'Select All Disease' should be chosen. The Word Cloud option can be turned off to increase the processing speed.

The Options section is shown below:



Another section is for Add-on Databases. In default, the 3 disease-gene mapping databases are selected: DisGenet, Genetic Association Database and Genecards database. The Mentha gnee-gene interaction database can also be turned on.

After all the input is set up, the 'Submit' button should be clicked to submit the job, and a page with the result URL immediately shows up.

Usually if there are only a few terms (less than 5), Phenolyzer's result will be available in 5 minutes (2 minutes without 'Word Cloud' selected).

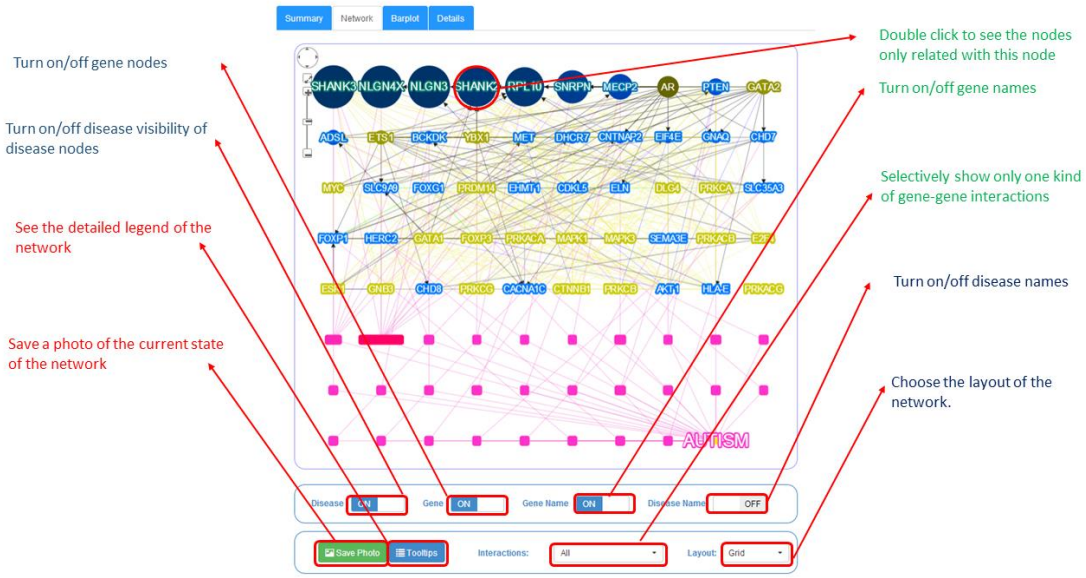
The result page consists of 4 sections: Summary, Network, Barplot and Details.
In the Summary section, the submission information is reported, like whether 'Phenotypes are interpreted', whether 'All diseases are considered', how many

genes are in the input gene list and how many disease terms are in the input and how many have corresponding records.

Below these brief summaries, the output files are available for download: the disease files can be got by clicking each disease term, the HPO files can be clicking the '(HPO)' after the disease, also the wordcloud image can be got by clicking the wordcloud tag after each term. The WHOLE REPORT and FINAL GENE LIST are actually the detailed report and the gene list with normalized scores mentioned for the output. If there is a gene list or a region file in the user's input, an additional report and gene list prioritizing these genes are also available. Beyond this, the detailed report and the normalized scores for the seed genes without gene network growth is also available.

The Detail section just presents part of the information of the detailed report, and the Barplot section just presents the top 500 genes with scores in the output list.

The Network instruction is shown below:



Command line tool instruction

Phenolyzer is open source and we believe it can be easily integrated into other gene/variant discovery software.
The Phenolyzer command line tool can be downloaded at
https://github.com/WangGenomicsLab/phenolyzer

You can either download the zip file by visiting this website or use Git pull command directly in your command line environment (no need to visit this website then):

```
git pull https://github.com/WangGenomicsLab/phenolyzer
```

NOTICE: This part can be only run in a Linux operation system (Ubuntu, centOS and so on).

After download, please extract in a Linux system, with command:

```
tar xvfz Phenolyzer.1.0.5.tar.gz
```

After extraction, get into its root directory:

```
cd command_line
```

Then try some examples:

Help:

```
perl disease_annotation.pl –help
```

Prioritize 'Sleep' genes with wordcloud:

```
perl disease_annotation.pl sleep -p -ph -logistic –wordcloud -out sleep/out
```

Use the terms in 'disease' file:

```
perl disease_annotation.pl disease -f -p -ph -logistic -out disease/out
```

Use the cnv.bed region and prioritize 'alzheimer' genes:

```
perl disease_annotation.pl alzheimer -bedfile cnv.bed -p -ph -logistic -out alzheimer/out
```

Use the Mentha gene-gene interaction database as Addon:

```
perl disease_annotation.pl alzheimer -p -ph -logistic -out alzheimer/out -addon_gg DB_MENTHA_GENE_GENE_INTERACTION -addon_gg_weight 0.05
```

Warning: Please output your result in a different folder, Phenolyzer will try to merge all the files with the name 'out_{disease}_gene_scores' and the old files might also be merged and affect your results!

The output files corresponding to 'autism' includes:

```
out.annotated_gene_list
out.annotated_gene_scores
out_autism_diseases
out_autism_gene_scores
out_autism_hpo
out_autism_wordcloud
out_autism_wordcloud.png
out.final_gene_list
out.merge_gene_scores
out.predicted_gene_scores
out_Rwordcloud.log
out.seed_gene_list
```

The out_autism_diseases and out_autism_hpo are the disease details and HPO file for the term autism.

The out_autism_wordcloud is the frequency count for the out_autism_diseases and out_autism_wordcloud.png is the actual word cloud image.

The out.merge_gene_scores and out.seed_gene_list are the detailed report and gene list for seed genes without growth.

The out.final_gene_list and out.predicted_gene_scores are the detailed report and gene list for all genes.
The out.annotated_gene_scores and out.annotated_gene_list are the detailed report and gene list for the genes confined by the input genelist or regions.

Integration with wANNOVAR/ANNOVAR

To integrate with wANNOVAR, please visit http://wannovar.usc.edu
You can either enter your phenotype/disease descriptions there. At the same time, you can save the wannovar '*_multianno.txt' file into your local computer, and use it as an input to the 'calculate_score.pl' script, instead of installing a local ANNOVAR yourself. The difference between the wANNOVAR online version and our 'calculate_score.pl' script is that 'calculate_score.pl' utilizes our own ensemble variant prioritization score called 'MetaSVM', which is published in a Human Molecular Genetics paper.

To integrate with ANNOVAR, please go to ANNOVAR's website at http://www.openbioinformatics.org/annovar/. After download, please extract the file:

```
tar xvfz annovar.latest.tar.gz
```

After extraction, enter the ANNOVAR directory:

```
cd annovar
```

Then download some necessary databases:

```
perl disease_annotation.pl -downdb -webfrom annovar -buildver hg19 refGene
/humandb

perl disease_annotation.pl -downdb -webfrom annovar -buildver hg19
ljb23_metasvm /humandb
```

Then convert your vcf file:

```
perl convert2annovar.pl -format vcf4 <vcf_file_name.vcf>  >  variant
```

IMPORTANT: Please use table_annovar.pl to annotate your vcf file:

```
perl table_annovar.pl <vcf_file_name.vcf> /humandb -outfile final -buildver
hg19  -protocol refGene,ljb23_metasvm  -operation g,f  -otherinfo
```

After this is finished, we need to create a new folder:

```
mkdir result
```

And copy Phenolyzer's result file and ANNOVAR's annotation file into it:

```
cp <phenolyzer_result_directory>/out.final.gene_list ./result
cp <annovar_result_directory>/final.hg19_multianno.txt ./result
```

Then please run our script provided with Phenolyzer:

```
cd result
perl <phenolyzer_directory>/calculate_score.pl out.final.gene_list
final.hg19_multianno.txt  > outfile.txt
```

And the 'outfile.txt' contains the prioritized variant list.