HPO富集分析的一些参考网站和包

> 往届学长作业：https://github.com/tongliu-liu/HPO-enrichment-analysis
>
> R包：HPOSim：https://pubmed.ncbi.nlm.nih.gov/25664462/#:~:text=The%20Human%20Phenotype%20Ontology%20%28HPO%29%20provides%20a%20standardized,used%20offline%20and%20provide%20only%20few%20similarity%20measures.
>
> **python HPO基因集合富集分析示例：**https://nanguage.github.io/examples/hpo_enrich/example_sagd_00055.html

> 官网，下载HPO数据，查HPO编号和注释用：https://hpo.jax.org/app/
>
> SAGD示例数据库（配合Python包食用）：http://bioinfo.life.hust.edu.cn/SAGD#!/download
>
> 人类疾病数据库（配合R包）：https://omim.org/

老师推荐：HPOterms在文本中得到，与水稻的难度类似?

挑战性：实验设计，获取基因集合，要把研究的问题立起来，要有基因的具体来源。

使用python包进行HPO富集分析

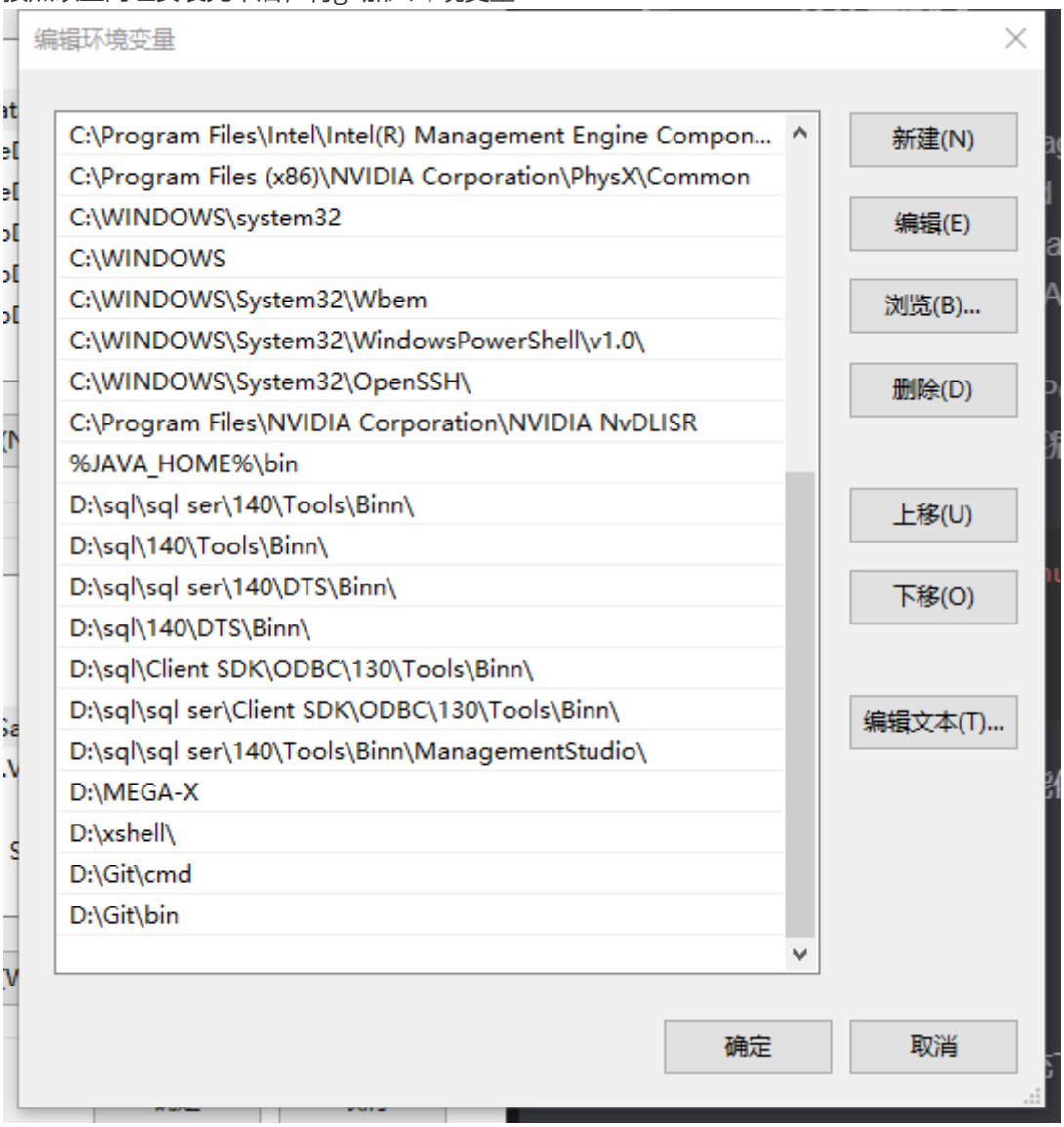
> 环境：Windows10下的vscode
>
> python 3.7.4 64-bit based on conda

依赖包的安装

首先要安装git（windows10系统下）

> 参考网址：https://git-scm.com/download/win
>
> https://blog.csdn.net/qq_32786873/article/details/80570783

按照以上网址安装完毕后，将git加入环境变量：



> 依赖包网址：https://github.com/Nanguage/BioTMCourse/tree/master/HPO%20enrich

```
git clone https://github.com/Nanguage/BioTMCourse.git
cd  D:\linuxvs\ziranyuyan\xiangmu\ceshi\ceshi\BioTMCourse\HPO enrich
python setup.py install #报错了，找不到request包
pip install request -i http://pypi.douban.com/simple/ --trusted-host
pypi.douban.com#终端中安装request
pip install bokeh
#安装完后再运行python setup.py install仍然报相同错误，先不管，尝试运行后续代码
```

依赖包准备

```
import pandas as pd
import matplotlib.pyplot as plt
from bokeh.plotting import show
from bokeh.io import output_notebook
output_notebook()

from hpoea.enrich import GSEA
from hpoea.plot import LineagePlot, dot_plot#导入包
```

数据准备

> 本笔记仅展示代码的可行性，不再对输入数据和数据背景，研究价值和具体的结果分析做出介绍，需要结合具体情况进行分析。
>
> 在这个例子中，我们使用性别相关的基因作为输入例子，它从SAGD数据库下载。选择数据组SAGD_00055(人类下丘脑组织)作为输入。

SAGD_00055.csv格式如图，这是基因表达差异分析的结果

A	B	C	D	E	F	G	H	I
	baseMean	log2FoldCl	lfcSE	stat	pvalue	padj	FPKM_M	FPKM_F
ENSG0000	850.4825	8.188046	0.218044	37.55222	0	0	6.015774	0.020867
ENSG0000	908.7269	8.637194	0.215148	40.14538	0	0	10.23449	0.026688
ENSG0000	1493.629	8.147712	0.208643	39.05098	0	0	39.94822	0.15503
ENSG0000	602.9739	8.827457	0.252515	34.95813	9.74E-268	1.16E-263	3.580869	0.007935
ENSG0000	343.5949	7.913893	0.237037	33.38676	2.13E-244	2.03E-240	1.98426	0.008775

其中log2FoldChange是对差异倍数取log2的值，这个值为负则对应基因为下调基因（相对表达量低），反之则为上调基因。

> http://www.pinlue.com/article/2019/07/1303/489298644023.html

padj是矫正过后的p值，越小则差异表达越显著。

```
input_csv = "D:/linuxvs/ziranyuyan/xiangmu/ceshi/ceshi/SAGD_00055.csv"#读入数据
exp = pd.read_csv(input_csv)
exp.columns = ["ensembl_id"] + list(exp.columns)[1:]
print(exp.head(3))
'''         ensembl_id     baseMean  ...      FPKM_M    FPKM_F
0  ENSG00000012817   850.482453  ...    6.015774  0.020867
1  ENSG00000067048   908.726851  ...  10.234487  0.026688
2  ENSG00000129824  1493.628628  ...  39.948216  0.155030

[3 rows x 9 columns]'''
```

数据筛选

```
#选择具有显著差异表达的条件基因:padj<= 0.05
sig = exp[exp.padj <= 0.05]
test_genes = list(sig.ensembl_id)
len(test_genes)#54
```

共选择了54个基因，以下是基因列表：

```
for i in range(len(test_genes)//5 + 1):
    print(" ".join(test_genes[i*5:(i+1)*5]))
'''ENSG00000012817 ENSG00000067048 ENSG00000129824 ENSG00000114374
ENSG00000131002
ENSG00000198692 ENSG00000165246 ENSG00000183878 ENSG00000229807 ENSG00000233864
ENSG00000099725 ENSG00000067646 ENSG00000099715 ENSG00000176728 ENSG00000154620
ENSG00000260197 ENSG00000206159 ENSG00000241859 ENSG00000278847 ENSG00000273906
ENSG00000227289 ENSG00000228764 ENSG00000232226 ENSG00000229236 ENSG00000169953
ENSG00000225117 ENSG00000270641 ENSG00000092377 ENSG00000267793 ENSG00000232348
ENSG00000224060 ENSG00000229163 ENSG00000215580 ENSG00000002586 ENSG00000231535
ENSG00000227494 ENSG00000133048 ENSG00000258484 ENSG00000184895 ENSG00000233070
ENSG00000228787 ENSG00000005889 ENSG00000214717 ENSG00000135245 ENSG00000064886
ENSG00000215301 ENSG00000229238 ENSG00000147050 ENSG00000198535 ENSG00000130600
ENSG00000217896 ENSG00000126012 ENSG00000261600 ENSG00000169248'''
```

格式转化

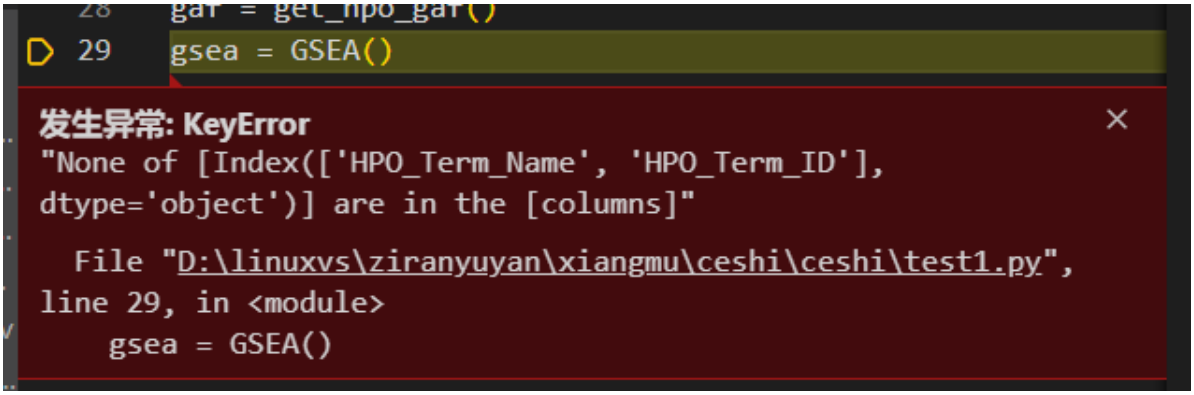
基因名称是ENS格式，但HPO需要Entrez格式。

把基因转换为正确的格式

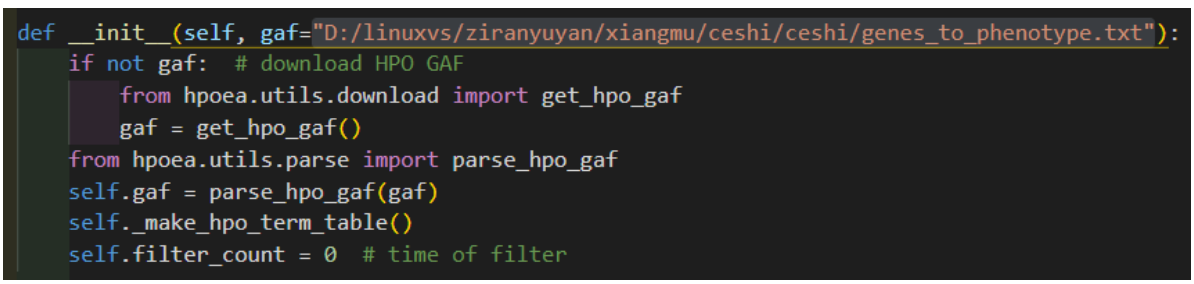
```
from hpoea.utils.idconvert import EntrezEnsemblConvert
cvt = EntrezEnsemblConvert() #报错，缺了个包，在终端中： pip3 install
biothings_client解决
test_entrez = cvt.ensembl2entrez(test_genes)
print(len(test_entrez))#36
#转换格式后缺失了一些基因（原本有54个）
```

富集分析

这里使用GSEA类中的一部分函数功能完成富集分析



在导入“类”GSEA时，出现以上报错，浏览enrich.py后发现是由于该函数自动下载的HPO数据格式出错导致，手动修改函数导入自己下载并修改为正确格式后的HPO文件，修改位置如图
C:\Users\Yangxiaolong\AppData\Local\Programs\Python\Python37\Lib\site-packages\hpoea-0.0.0-py3.7.egg\hpoea



完成以上错误处理后开始进行富集分析：

```
gsea = GSEA()
gsea.enrich(test_entrez) #富集分析
gsea.multiple_test_corretion(method='fdr_bh')
print(gsea.enrichment_table.head(1)) #查看结果的第一行
```

结果第一行如下

```
   HPO_term_ID         HPO_term_name  gene_num  study_count  n_study  ...  background_ratio   odd_ratio        pvalue          padj                                  related_g
enes
55  HP:0001450  Y-linked inheritance        17            5      512  ...         0.000073  134.489315  2.970414e-10  8.940945e-08  KDM5D USP9Y DDX3Y PRY2 TBL1Y RPS4Y2 PRY BPY2
V...
```

HPOID：即对应概念基因集在HPO数据库的ID

HPONAME：简单的注释（注释的题目），在官网搜索对应的ID可以看到更详细的注释内容

relatedgene：当前HPO基因集下与输入的基因集有关联的基因

pvalue：反映了输入基因集与对应HPO基因集及其概念的关联程度，关联越密切，p值越小

padj：p的矫正值

```
#gsea.enrichment_table.shape[0]
gsea.filter(by='padj', threshold=0.05)#筛选padj小于0.05
#print(type(gsea.enrichment_table))
t=gsea.enrichment_table
t.to_csv("enrichiment_table.csv")#保存分析结果表格为enrichiment_table.csv
```

结果中的前五行如图所示（共有54行）：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		HPO_term	HPO_term	gene_num	study_cour	n_study	population	n_populati	gene_ratio	backgroun	odd_ratio	pvalue	padj	related_genes			
2	55	HP:000145	Y-linked in	17	5	512	17	234119	0.009766	7.26E-05	134.4893	2.97E-10	8.94E-08	KDM5D USP9Y DDX3Y PRY2 TBL1Y RPS4Y2			
3	56	HP:000002	Azoosperm	105	6	512	105	234119	0.011719	0.000448	26.12935	1.42E-07	2.14E-05	KDM5D USP9Y NR0B1 NR0B1 ANK1 DHX37			
4	245	HP:000765	Eversion of	10	3	512	10	234119	0.005859	4.27E-05	137.1791	1.23E-06	0.000124	CDC42 CDC42 GRIA3 HNRNPK HNRNPK KD			
5	13	HP:000873	Decreased	209	6	512	209	234119	0.011719	0.000893	13.12719	7.86E-06	0.000592	DNAJC19 KDM5C KDM5C USP9Y NR0B1 NF			
6	182	HP:010077	Urogenital	27	3	512	27	234119	0.005859	0.000115	50.80707	2.92E-05	0.001558	MKKS NR0B1 DHX37 DHX37 CHRM3 DMRT			

由于padj<=0.05的结果太多，为了方便展示，筛选padj<0.01的结果

```
gsea.filter(by='padj', threshold=0.01)#筛选padj小于0.01
t.to_csv("enrichiment_table0.01.csv")#保存分析结果表格为enrichiment_table0.01.csv
```

得到如下图18行结果：

	A	B	C	D	
1		HPO_term_	HPO_term_	gene_num	stu
2	55	HP:000145	Y-linked in	17	
3	56	HP:000002	Azoosperr	105	
4	245	HP:000765	Eversion of	10	
5	13	HP:000873	Decreased	209	
6	182	HP:010077	Urogenital	27	
7	237	HP:000923	Short 5th f	28	
8	100	HP:000266	Nephrobla	83	
9	151	HP:001196	Elevated ci	34	
10	150	HP:000823	Elevated ci	38	
11	156	HP:001286	Ovotestis	7	
12	58	HP:000325	Male infert	120	
13	144	HP:000199	Neonatal h	51	
14	178	HP:000168	Coarctatio	131	
15	232	HP:000473	Crossed fu	11	
16	188	HP:000015	Gonadobla	13	
17	159	HP:000014	Polycystic	63	
18	34	HP:000002	Cryptorchi	1010	
19	91	HP:000142	X-linked d	72	

结果可视化

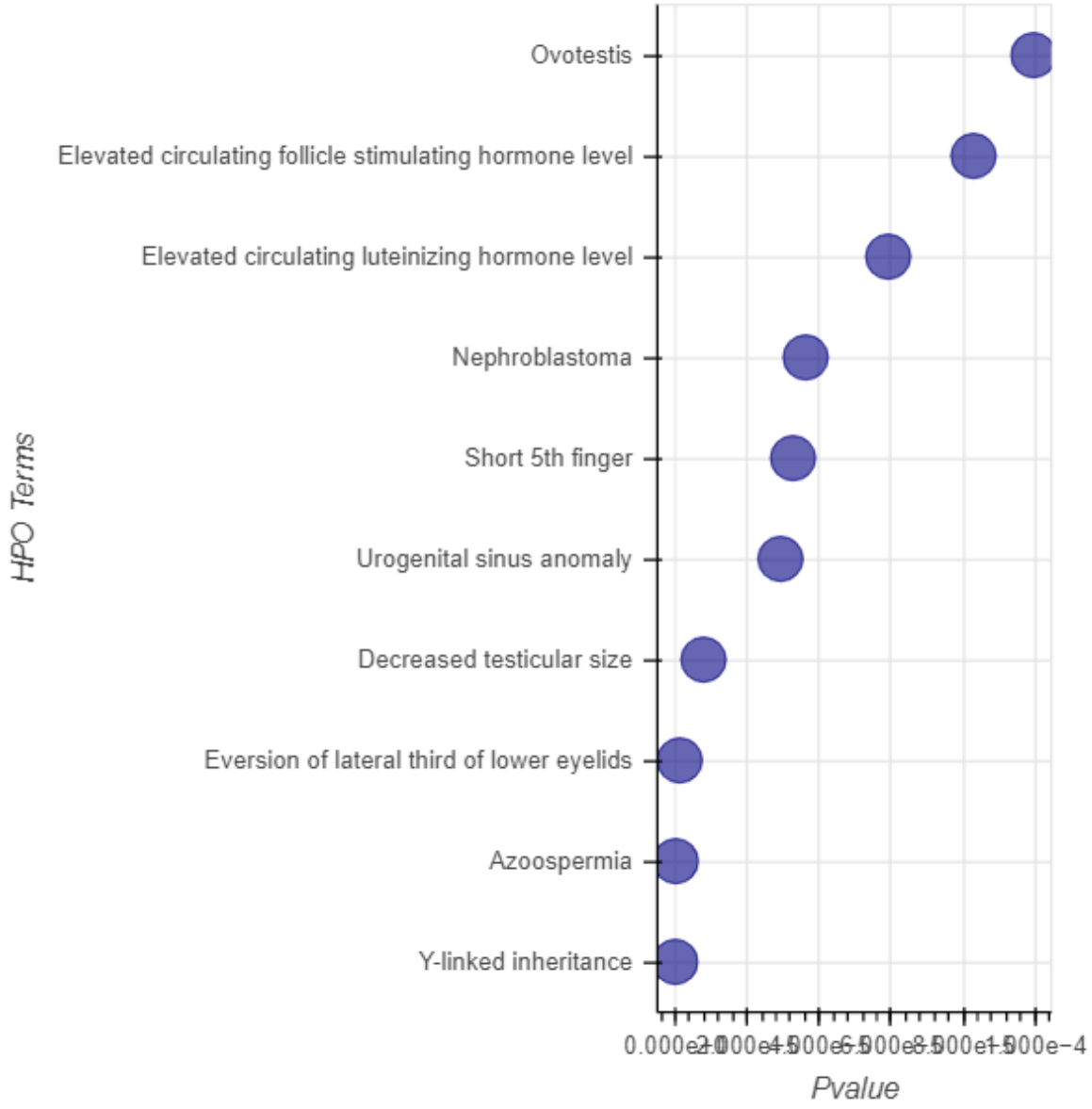
绘图函数dotplot.py：

```
import pandas as pd
import matplotlib.pyplot as plt
from bokeh.plotting import show
from bokeh.io import output_notebook
output_notebook()
from hpoea.enrich import GSEA
from hpoea.plot import LineagePlot, dot_plot
def dotp(path):
    f = open(path, encoding='utf-8')
    data = pd.read_csv(f)
    p = dot_plot(data, size=20, x='pvalue')
    return p
```

对应的主函数main.py：

```
import dotplot
from PIL import Image
import matplotlib.pyplot as plt
from bokeh.plotting import show
p=dotplot.dotp("D:/linuxvs/enrichiment_table0.01.csv")
#p.savefig("D:/linuxvs/dot.png")
show(p)
#print(type(p))
```

结果如下：



绘制网络图：

绘图函数mapplot.py

```
import pandas as pd
import matplotlib.pyplot as plt
from bokeh.plotting import show
from bokeh.io import output_notebook
output_notebook()
from hpoea.enrich import GSEA
from hpoea.plot import LineagePlot, dot_plot
def mapp(path):
    f = open(path, encoding='utf-8')
    data = pd.read_csv(f)
    terms = list(data.HPO_term_ID)
    lin = LineagePlot()
    fig, ax = plt.subplots(figsize=(20, 10))
    lin.plot(terms, ax=ax)
```

对应主函数部分：

```
mapplot.mapp("D:/linuxvs/enrichiment_table0.01.csv")
```

运行时以上函数时报错，经过排查，是由于撞墙而不能下载绘制网络图所需文件导致

> 手动下载：https://raw.githubusercontent.com/obophenotype/human-phenotype-ontology/master/hp.obo
>
> 保存为hpo.obo

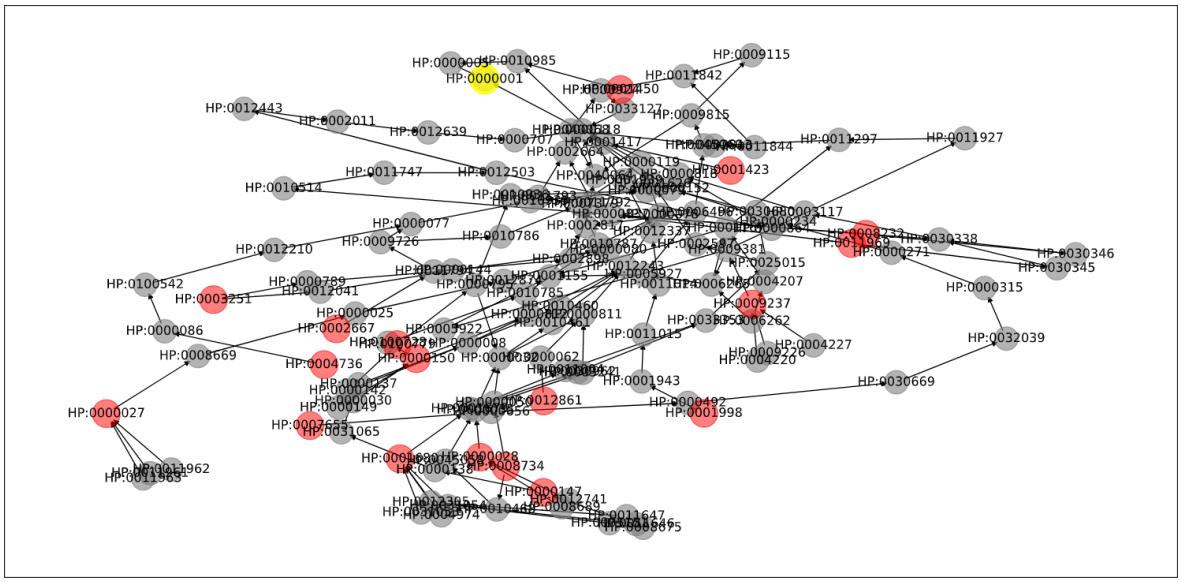
修改plot.py中的如图部分，导入自己下载的文件

```
class LineagePlot(object):
    """Plot the tree structure of OBO file
    """
    def __init__(self, obo_file="D:/linuxvs/ziranyuyan/xiangmu/hpo.obo"):
        if not obo_file:  # download HPO OBO file
            from hpoea.utils.download import get_hpo_obo
            obo_file = get_hpo_obo()
        from hpoea.utils.parse import parse_hpo_obo
        self.obo = parse_hpo_obo(obo_file)
```

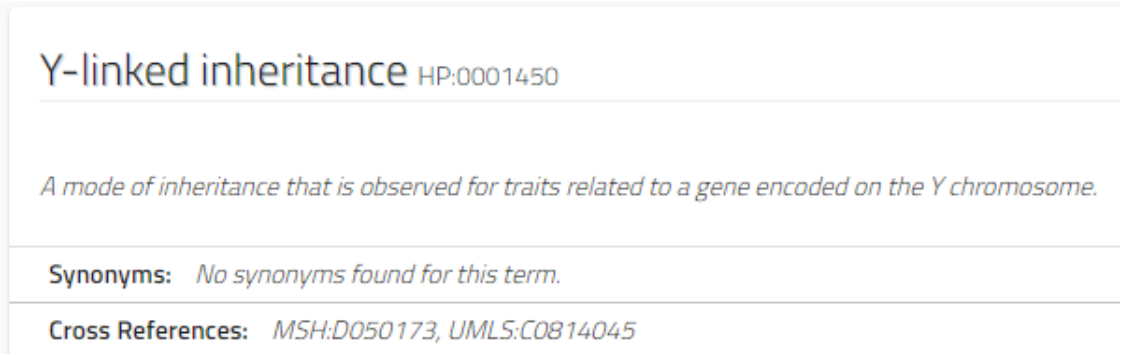
再执行提示缺少部分依赖包：

```
pip install graphviz
pip install pygraphviz
```

解决以上问题再运行主函数，获得图片如下：



> 代码到此为之运行完毕。
>
> 进一步分析贼需要在官网查询更加详细的注释并进行推理分析，本笔记仅作举例：



HP0001450注释信息：一种与Y染色体编码的基因相关的遗传模式。

由此可以推理输入的基因集与Y染色体可能有关，而实际上我们输入的基因集就是下丘脑组织中与性别相关的基因，因此分析较为合理。

R包HPOSim的使用测试（可能是一种更深度，更有价值的HPO分析）：

> HPOSim is an R package for analyzing phenotypic similarity for genes and diseases based on HPO data. Seven commonly used semantic similarity measures are implemented in HPOSim. Enrichment analysis of gene sets and disease sets are also implemented, including hypergeometric enrichment analysis and network ontology analysis (NOA).
>
> HPOSim是一个基于HPO数据分析基因和疾病表型相似性的R包。在HPOSim中实现了七种常用的语义相似性度量。实现了基因集和疾病集的富集分析，包括超几何富集分析和网络本体分析。
>
> HPOSim consists of two parts: (i) the similarity measures between phenotypes (HPO terms), between human genes (Entrez IDs) and between diseases (OMIM IDs), and (ii) HPO-based enrichment analysis (NOA and the hypergeometric method) for gene set and disease set.
>
> HPOSim包括两个部分:(i)表型之间(HPO)、人类基因之间(Entrez id)和疾病之间(OMIM id)的相似性度量，以及(ii)基于HPO的基因集和疾病集的富集分析(NOA和超几何方法)。
>
> 结果示例：
>
> Table 2
> **Gene modules of the aging network.**
>
> | Module | Size | Genes (Entrez ID) | TOP 5 Enriched GO BP Terms | TOP 5 Enriched HPO Terms | TOP 5 Enriched KEGG Pathways |
> |---|---|---|---|---|---|
> | M1 | 36 | 25, 207, 472, 581, 596, 641, 672, 675, 701, 1029, 1050, 1499, 1956, 2064, 2308, 3265, 4193, 4292, 4609, 5159, 5422, 5728, 5781, 5925, 6794, 7015, 7157, 7486, 9184, 1385, 7153, 627, 1649 | regulation of apoptosis, cell cycle process, regulation of programmed cell death, regulation of cell death, regulation of cell cycle | Neoplasm, Neoplasm by anatomical site, Neoplasm by histology, Sarcoma, Hematological neoplasm | Pathways in cancer, Prostate cancer, Endometrial cancer, Glioma, Bladder cancer |
> | M2 | 26 | 545, 1387, 2010, 2033, 2068, 2073, 2074, 2260, 3479, 3480, 4000, 4036, 4792, 4803, 5979, 7020, 7314, 7341, 7415, 7507, 5830, 1950, 1161, 847, 1490, 2067 | DNA metabolic process, response to UV, response to radiation, DNA repair, nucleotide-excision repair | Intrauterine growth retardation, Aplasia/Hypoplasia of the mandible, Micrognathia, Defective DNA repair after ultraviolet radiation damage, Abnormality of the mandible | Nucleotide excision repair, Prostate cancer, Pathways in cancer, Melanoma, Adherens junction |
> | M3 | 17 | 367, 2099, 2353, 2690, 2908, 3630, 3643, 3952, 3953, 5449, 5578, 6777, 7040, 8626, 8820, 2688, 5626 | response to hormone stimulus, response to endogenous stimulus, response to organic substance, positive regulation of macromolecule metabolic process, response to estrogen stimulus | Abnormality of the anterior pituitary, Abnormality of the pituitary gland, Abnormality of the endocrine system, Abnormality of the hypothalamus-pituitary axis, Anterior hypopituitarism | Jak-STAT signaling pathway, Neuroactive ligand-receptor interaction, Cytokine-cytokine receptor interaction, Aldosterone-regulated sodium reabsorption, Pathways in cancer |
> | M4 | 11 | 355, 2071, 3561, 3575, 4683, 4791, 5295, 5580, 6774, 6929, 5336 | cell activation, B cell activation, lymphocyte activation, leukocyte activation, immune system development | Abnormality of lymphocytes, Abnormal immunoglobulin level, Abnormality of B cell physiology, Abnormality of B cells, Abnormality of humoral immunity | Pathways in cancer, Jak-STAT signaling pathway, Fc epsilon RI signaling pathway, Fc gamma R-mediated phagocytosis, Neurotrophin signaling pathway |
> | M5 | 9 | 3064, 4001, 4137, 5155, 6872, 6908, 5663, 6647, 1938 | negative regulation of neuron apoptosis, regulation of neuron apoptosis, positive regulation of MAP kinase activity, behavior, regulation of membrane potential | Abnormality of extrapyramidal motor function, Personality changes, Adult onset, Dysarthria, Parkinsonism | Huntington's disease, Basal transcription factors |
> | M6 | 5 | 348, 351, 3717, 2876, 5328 | regulation of response to external stimulus, induction of apoptosis, induction of programmed cell death, positive regulation of apoptosis, positive regulation of programmed cell death | Long-tract signs, Abnormal bleeding, Abnormalities of the peripheral arteries, Arterial stenosis, Cerebral inclusion bodies | N/A[a] |

```
install.packages("D:/linuxvs/ziranyuyan/xiangmu/ceshi/HPO.db_1.9.tar.gz", repos = NULL, type = "source")
```

我没有找到这个包的代码示例，只能依靠论文和help（）文件获取帮助，进度缓慢，不再赘述。