HPO富集分析课程论文草稿

生信1802杨晓龙

2018317220204

Abstact

大脑使用多巴胺来发送信息和协调控制运动，包括从走路到说话，写作甚至微笑都与其相关。当大脑中产生多巴胺的细胞出现问题时，帕金森病就产生了。我们目前并不清楚多巴胺细胞丢失的原因，但是研究人员们正在努力寻找保护这些细胞的方法。我们的研究使用了HPO富集分析了从NCBI下载的人类帕金森相关基因，并试图从结果中找到HPO注释与帕金森的联系，从而使我们更加了解这种疾病。（再从知网或者sciencedirect找一两篇论文稍微修改添加，引用一下把）

关键词：帕金森，HPO富集分析，表型特征

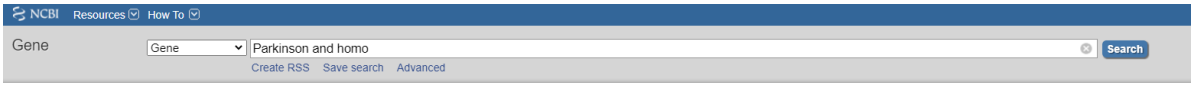
课题概况（项目设计）：帕金森病

选修本课程的目的是根据以往上过本课的同学推荐，此课学习的内容全部都是生信较为前沿，且贴合实际工作的内容，老师授课水平也比较高，且课程通过率较高，没有期末考试。

选择本课题是因为其难度较为适中，且HPO的注释集相比GO，其concept更加宽泛，尽管分析过程相比已经十分成熟的GO分析可能更加复杂，但是其结果可能更易于理解。选择此课题的原因是因为家中有老人疑似得了此病，我也曾搜索过相关科普视频进行了解，了解到目前并没有根治帕金森的方法，因此想借着本次机会自己动手完成此次针对它的HPO分析，了解帕金森表型特征以及与HPO中其他表型异常间的关系，并获取相关的基因集表格，加深对此病的理解并且为进一步分的分析做准备。

数据来源以及数据格式介绍

首先在NCBI Gene数据库下载关于帕金森病的相关symbol资料：



提取下载得到的symbol列，即为基因编号，保存为unicode（utf-8）格式，并删去题头得到文件symbol_Parkinson.txt，共有1列，345行。

人类表型本体论(HPO)提供了在人类疾病中遇到的表型异常的标准化词汇。HPO中的每个术语都描述了一种表型异常。本文的分析中用到了HPO的两类数据，分别是：

hpo.obo：用于构建网络图数据文件，从HPO官网下载。格式说明见以下网址：https://www.cnblogs.com/wangshicheng/p/10107463.html

```
[Term]
id: HP:0000002
name: Abnormality of body height
def: "Deviation from the norm of height with respect to that which is expected according to age and gender norms." [HPO:probinson]
synonym: "Abnormality of body height" EXACT layperson []
xref: UMLS:C4025901
is_a: HP:0001507 ! Growth abnormality
created_by: peter
creation_date: 2008-02-27T02:20:00Z
```

phenotype_to_genes.txt：用于匹配HPOterms的基础文件，从HPO官网下载。

```
#Format: HPO-id<tab>HPO label<tab>entrez-gene-id<tab>entrez-gene-symbol<tab>Additional Info from G-D source<tab>G-D source<tab>disease-ID for link
HP:0000002	Abnormality of body height	5073	PARN	orphadata	ORPHA:1775
HP:0000002	Abnormality of body height	10084	PQBP1	orphadata	ORPHA:93950
HP:0000002	Abnormality of body height	50485	SMARCAL1	-	mim2gene	OMIM:242900
HP:0000002	Abnormality of body height	2317	FLNB	orphadata	ORPHA:503
```

> 所有代码中使用的python包和部分解释参考我的另一篇笔记：HPO富集分析项目笔记.md，请把两个笔记对照着观看，用的包完全一致，重复内容这里不再赘述。

研究方法

使用https://nanguage.github.io/examples/hpo_enrich/example_sagd_00055.html以及其相关网站获取的python包，根据本次项目的要求略作修改，具体修改的内容在代码实践部分再做介绍。

算法背景

> 参考网址：什么是富集分析https://www.zhihu.com/question/30778984
>
> p值计算：fisher精确分析的实现：https://docs.scipy.org/doc/scipy-0.17.0/reference/generated/scipy.stats.fisher_exact.html
>
> 采用BH法校正多重检验的p值：https://blog.csdn.net/zhu_si_tao/article/details/71077703（自行查阅多重检验矫正的BH法进行补充说明）
>
> 与GO分析的联系与区别（自行查阅课程项目3GO分析的相关源代码进行对比）

核心思路

1，从NCBI上下载symbol数据，并自行构建python函数完成symbol编号与ENSG编号的转换，方便使用工具包。

2，调试并适当修改调整工具包的代码，使其能够完成本项目的HPO分析工作，获取分析结果表格。

3，根据分析结果绘制泡泡图与网络图。

4，对结果进行详尽注释的检索和进一步的生物学分析。

本文与课堂内容的区别或者补充

课堂上并没有讲授HPO分析的具体代码实现，本文在python上实现了类似于课堂讲授的GO富集分析的过程，且获得了类似的分析结果。（可以结合项目3GO分析极其源代码自行补充相似与联系）

代码操作

> 环境:WIN10下的VSCODE + Powershell终端+Python3.7.4 64bit

任务描述

根据已有的python包，进行适当修改，增添代码后完成从NCBI下载目标Gene Symbol并完成HPO分析的流程。最终得到分析结果的csv表格和泡泡图与网络图。

实验设计

1，数据预处理请见本文的“数据来源以及数据格式介绍”部分。

2，数据集的划分（输入数据格式）：与人类帕金森相关的基因的整个symbo列都作为输入数据。

3，算法实施：由于fisher检验和BH矫正方法都已经非常成熟，本文算法部分都采用了现成的python包，没有自行编写新算法。

4，硬件配置：intel i5 8代标压处理器。环境： Windows10家庭版下的Vscode + Powershell终端 +Python3.7.4 64bit

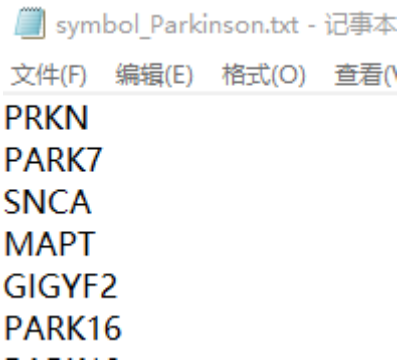
转换symbol为ENSG编号（HPO富集分析项目笔记.md中没有）

程序StoENSG.py

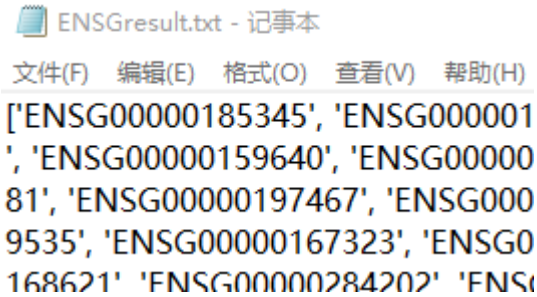
依赖于mygene包，读取symbol文件转换为对应的ENSG编号并生成文件ENSGresult.txt

```
#代码来源：生信1802杨晓龙 原创
#2021/5/12
#将symbol编号转化为ENSG编号
import mygene
mg = mygene.MyGeneInfo()

symb = []
with open("symbol_Parkinson.txt") as f: #读入symbol文件
    for l in f:
        symb.append(l.strip())
#print(symb)
out = mg.querymany(symb, scopes='symbol', fields='ensembl.gene',
species='human')
#print(out)

result = []
for i in out:                       #获取对应的ENSG编号列表
    if 'ensembl' in i.keys():
        genelist = i['ensembl']
        if type(genelist) == list:
            for j in genelist:
                result.append(j['gene'])
        else :
            result.append(genelist['gene'])

with open("ENSGresult.txt", "w") as output: #保存为txt文件
    output.write(str(result))
```

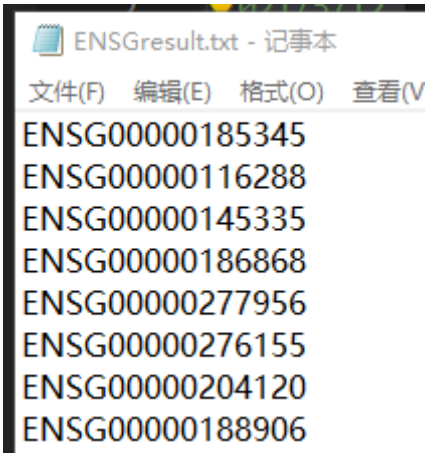
输入文件：



输出文件：



输出文件是一个字符串列表，使用excel处理，删去多余的符号并转为一列



修改后的ENSGresult.txt如上图所示,有433行。

HPO富集分析

程序HPOEEnrich.py

```
#代码来源：生信1802杨晓龙 原创
#2021/5/12
#这个函数能生成main函数需要的前置文件
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

input_txt = "D:/linuxvs/ziranyuyan/xiangmu/ceshi/ceshi/ENSGresult.txt"
ENSG=[]
with open(input_txt) as f: #读入symbol文件
    for l in f:
```

```
        ENSG.append(l.strip())
print(ENSG[0:3])

for i in range(len(ENSG)//5 + 1):
    print(" ".join(ENSG[i*5:(i+1)*5]))

from hpoea.utils.idconvert import EntrezEnsemblConvert
cvt = EntrezEnsemblConvert()
ENTRZ = cvt.ensembl2entrez(ENSG)
print(ENTRZ[0:3])
print(len(ENTRZ))

gsea = GSEA()
gsea.enrich(ENTRZ)
gsea.multiple_test_corretion(method='fdr_bh') #采用BH法矫正p值
print(gsea.enrichment_table.head(1))
gsea.enrichment_table.shape[0]
gsea.filter(by='padj', threshold=0.01)#筛选padj小于0.01
#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
```

以上步骤gsea.enrich(ENTRZ)中的核心代码：

```
#代码来源：https://github.com/Nanguage/BioTMCourse/tree/master/HPO%20enrich
        for term_id in tqdm(possi_terms):  # calculate the p-value of each
possible terms
            genes = list(self.gaf[self.gaf.HPO_Term_ID ==
term_id].entrez_gene_symbol) #提取具有相同id的所有term方便后续计算
            related_genes.append(genes)
            counts = self._get_counts(term_id)
            all_counts.append(counts)
            pval = self._calc_pvalue(*counts)
            pvals_uncorr.append(pval)
        study_count, n_study, population_count, n_population = zip(*all_counts)
```

绘制网络图,并控制网络的大小

函数mapplot.py

```
#代码来源：生信1802杨晓龙 原创
#2021/5/12
#这是网络图绘制函数
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)
```

绘制点图

```
#代码来源：生信1802杨晓龙 原创
#2021/5/12
#这是点图绘制函数
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)
    print(data)
    p = dot_plot(data, size=20, x='pvalue')
    return p
```

主函数,调用以上所有程序的结果生成点图和网络图

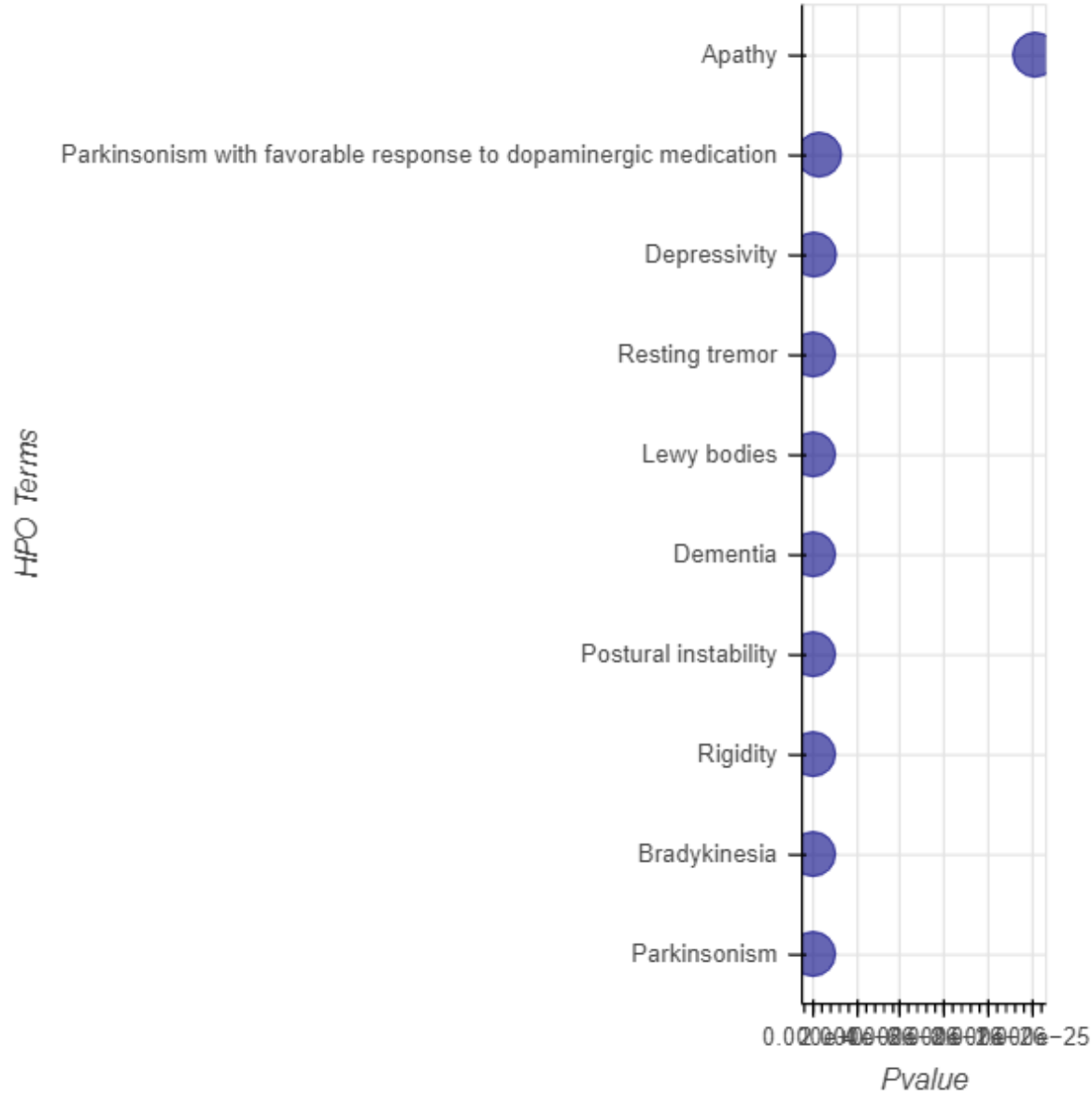
```
#代码来源：生信1802杨晓龙 原创
#2021/5/12
#主函数
import dotplot
from PIL import Image
import matplotlib.pyplot as plt
from bokeh.plotting import show
import mapplot
p=dotplot.dotp("D:/linuxvs/ziranyuyan/xiangmu/ceshi/ceshi/enrichiment_table.csv"
)
#p.savefig("D:/linuxvs/dot.png")
show(p)
#print(type(p))
mapplot.mapp("D:/linuxvs/ziranyuyan/xiangmu/ceshi/ceshi/enrichiment_table.csv")
```

结果展示

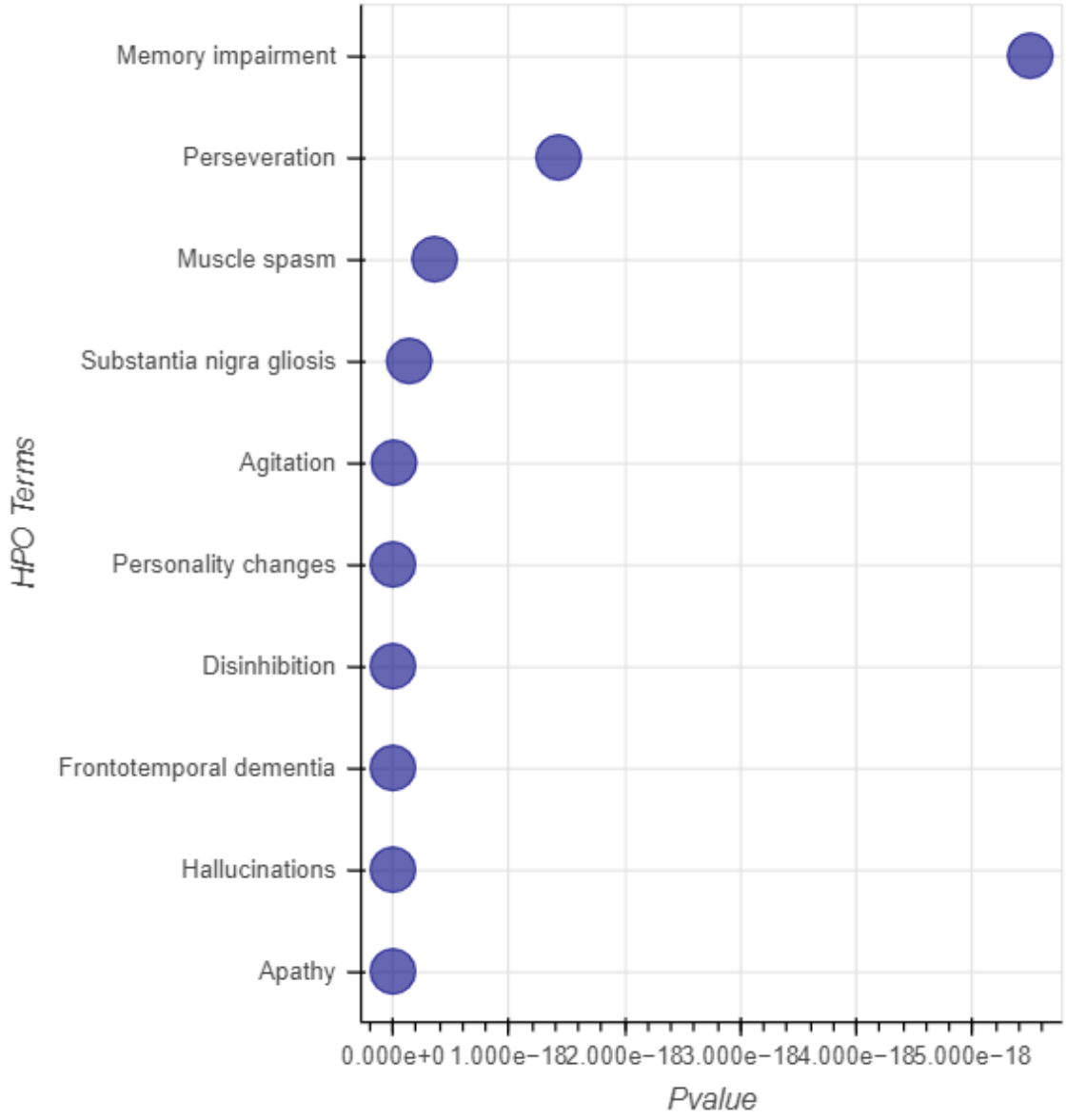
enrichiment_table.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1		HPO_term	HPO_term	gene_num	study_cour	n_study	population	n_populati	gene_ratio	backgroun	odd_ratio	pvalue	padj	related_genes													
2	2	HP:000130	Parkinsonis	155	74	10814	155	234119	0.006843	0.000662	10.33595	6.93E-56	1.65E-52	A2M MYORG ADH1C PCDH19 PLA2G6 ALS2 APOE APP APP PRKRA PRKRA ATP1A3 ATP1A3 SYNJ1 SQSTM1 CAT XPR1 LYST JAM2 CLN3 FBXO7													
3	105	HP:000206	Bradykines	162	75	10814	162	234119	0.006935	0.000692	10.02297	2.27E-55	2.70E-52	ACTA1 MYORG ADH1C PCDH19 PLA2G6 PLA2G6 PLA2G6 SLC25A4 PRKRA PRKRA PDE8B PDE8B ATP1A3 ATP1A3 SYNJ1 SYNJ1 SYNJ1 JAM2 FB													
4	118	HP:000206	Rigidity	203	74	10814	203	234119	0.006843	0.000867	7.89198	1.19E-45	9.43E-43	AARS1 ACTA1 ADAR ADH1C PLA2G6 PLA2G6 CASK PDE8B PDE8B ATP6V1A SYNGAP1 SYNJ1 SYNJ1 SYNJ1 SYNJ1 BRAT1 CACNA1A CACNA1B													
5	113	HP:000217	Postural in	106	50	10814	106	234119	0.004624	0.000453	10.21209	6.07E-38	3.62E-35	ADH1C PLA2G6 PLA2G6 ABCC6 ATP1A2 ATP1A3 ATP1A3 RNASEH1 SYNJ1 SYNJ1 SYNJ1 CACNA1A CACNA1A CACNB4 ERCC8 ERCC8 CLN5 FB													
6	0	HP:000072	Dementia	222	68	10814	222	234119	0.006288	0.000948	6.631415	1.48E-36	7.08E-34	A2M ADH1C AARS2 ABCD1 APOE APOE APOE APP APP APP APP APP GBA2 GBA2 ARSA ASAH1 ATP6V1A ATP6V1E1 CTSF ATP7B AUH SC													
7	111	HP:010031	Lewy bodi	31	28	10814	31	234119	0.002589	0.000132	19.5545	1.54E-34	6.10E-32	ADH1C PLA2G6 FBXO7 SNCAIP GIGYF2 C19ORF12 RAB39B EIF4G1 EIF4G1 GBA GBA GBA GLUD2 GRN MAPT NR4A2 VPS13C LRRK2 LRRK2 ATX													
8	102	HP:000232	Resting tre	54	33	10814	54	234119	0.003052	0.000231	13.23032	1.62E-30	5.51E-28	ADCY5 ADCY5 ADH1C SLC25A4 ATP1A3 SYNJ1 CACNA1G RRM2B SNCAIP GIGYF2 GIGYF2 DNAJC6 RAB39B DNMT1 ATP6AP2 ATP6AP2 EIF4G1													
9	107	HP:000071	Depressivit	483	88	10814	483	234119	0.008138	0.002063	3.944444	4.27E-28	1.27E-25	SMC1A ADH1C AARS2 PLA2G6 PLA2G6 KCNT1 KCNT1 ANG SLC25A4 PRSS12 ANXA11 AR ARSA ARVCF JRK ABCB11 ATP1A3 ATP1A3 STX16 C													
10	172	HP:000254	Parkinsonis	27	23	10814	27	234119	0.002127	0.000115	18.44227	2.76E-27	7.31E-25	PLA2G6 FBXO7 GIGYF2 GIGYF2 EIF4G1 GBA GCH1 PDE10A HTRA2 PARK7 MAPT MAPT MAPT POLG POLG LRRK2 LRRK2 VPS35 SNCA TAF1 TAF													
11	564	HP:000074	Apathy	140	45	10814	140	234119	0.004161	0.000598	6.958807	1.01E-25	2.41E-23	ACAT1 SMC1A SYNJ1 SQSTM1 SQSTM1 SQSTM1 FOXH1 FOXH1 FOXH1 FOXH1 CACNA1A ATXN10 CP CHMP2B CHMP2B GIGYF2 DNAJC6 DCT													
12	98	HP:000073	Hallucinati	144	45	10814	144	234119	0.004161	0.000615	6.765507	3.85E-25	8.34E-23	ADH1C APP ARSA ARSA ARSA ARSA BCKDHA BCKDHB BCS1L BMPR1A SYNJ1 SQSTM1 CHI3L1 CLN5 WHRN COMT CPOX SNCAIP CTSH CYP27													
13	170	HP:000214	Frontotem	50	28	10814	50	234119	0.002589	0.000214	12.12379	1.28E-24	2.54E-22	PLA2G6 PLA2G6 SQSTM1 SQSTM1 SQSTM1 CCNF CHMP2B CHMP2B CHMP2B CYLD DCTN1 FUS GRN GRN GRN HNRNPA1 HNRNPA2B1 HNRN													
14	362	HP:000073	Disinhibitio	43	25	10814	43	234119	0.002312	0.000184	12.58699	1.08E-22	1.97E-20	APP SQSTM1 SQSTM1 SQSTM1 SQSTM1 CHMP2B CHMP2B ABCA7 ABCA7 TOMM40 FMR1 FTL FUS GRN GRN HTT MAPT MAPT MAPT TBK1 TBK1 TREM													
15	115	HP:000075	Personality	78	32	10814	78	234119	0.002959	0.000333	8.881896	1.69E-22	2.88E-20	ADH1C PLA2G6 PLA2G6 APOE ATP7B SQSTM1 SQSTM1 EIF2B4 EIF2B3 EIF2B2 EIF2B5 MMACHC CHMP2B CHMP2B CHMP2B SNCAIP DCTN1 TIN													
16	289	HP:000071	Agitation	106	35	10814	106	234119	0.003237	0.000453	7.14846	8.73E-21	1.39E-18	ACAT1 ANG ANXA11 APP SYNJ1 SQSTM1 BRAT1 CFAP410 CACNA1A CCNF VAPB COX10 GABBR2 GABBR2 CHMP2B GIGYF2 MATR3 DAO DNA													
17	108	HP:001196	Substantia	16	15	10814	16	234119	0.001387	6.83E-05	20.29652	1.41E-19	2.10E-17	ADH1C FBXO7 SNCAIP GBA GLUD2 MAPT ATXN3 ATXN3 ATXN3 NR4A2 PRKN LRRK2 ATXN2 ATXN8OS SNCA TBP													
18	281	HP:000339	Muscle spa	171	42	10814	171	234119	0.003884	0.00073	5.317451	3.59E-19	5.03E-17	ACADM ACADVL AMPD1 AMPD3 ANG SLC25A4 ANXA11 AR AR ATP1A1 STX16 SYNJ1 SQSTM1 CFAP410 CACNA1S CRPPA CASQ1 CASQ1 CA													
19	589	HP:003022	Perseverat	34	20	10814	34	234119	0.001849	0.000145	12.73507	1.43E-18	1.89E-16	SQSTM1 SQSTM1 CHMP2B CHMP2B TACO1 NPC2 FUS GRN GRN GRN MAPT MAPT MAPT TBK1 NONO TREM2 TREM2 TMEM106B TMEM106B													
20	308	HP:000235	Memory in	150	38	10814	150	234119	0.003514	0.000641	5.484571	5.50E-18	6.90E-16	KLRC4 MYORG ABCD1 ABCD1 APOE APOE APP APP FAS ARSA BMPR1A SQSTM1 EIF2B4 EIF2B3 EIF2B2 EIF2B5 ACSF3 C4A XPR1 CCR1 CP MMA													
21	104	HP:000133	Dystonia	466	70	10814	466	234119	0.006473	0.00199	3.252089	6.44E-18	7.67E-16	ACADS DNAJC19 ACOX1 ADAR ADAR ADAR ADCY5 ADCY5 MYORG TBC1D24 TBC1D24 ADH1C AARS2 HACE1 HACE1 PLA2G6 PLA2G6 PLA2G													

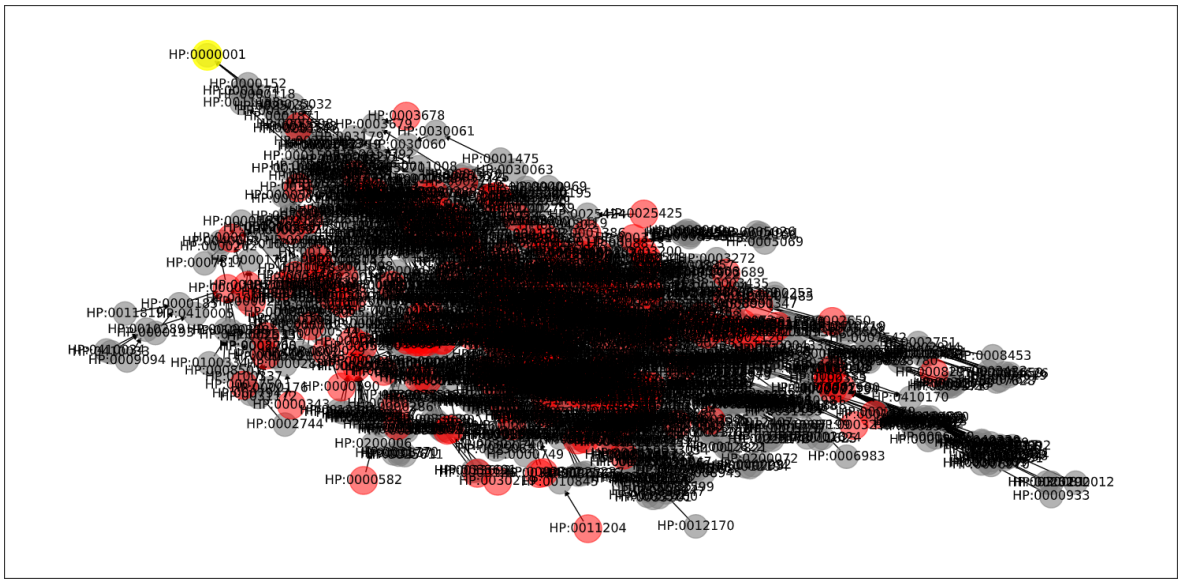
点图1（展示了前十个）：



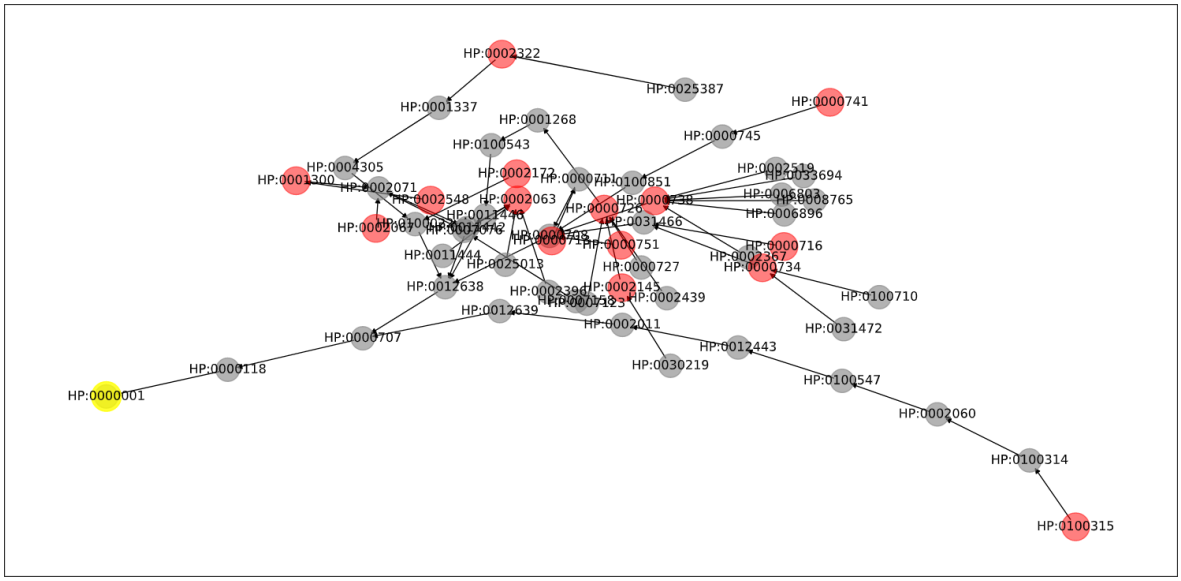
点图2（展示了第10到到19个）



网络图（不用贴这张）：



优化网络图(仅选取前15个点作图)



主要生物信息学实验和实验结论

p值最小的前20个term如图：

		HPO_term_	HPO_term_
2	2	HP:000130	Parkinsonis
3	105	HP:000206	Bradykines
4	118	HP:000206	Rigidity
5	113	HP:000217	Postural ins
6	0	HP:000072	Dementia
7	111	HP:010031	Lewy bodi
8	102	HP:000232	Resting tre
9	107	HP:000071	Depressivit
10	172	HP:000254	Parkinsonis
11	564	HP:000074	Apathy
12	98	HP:000073	Hallucinati
13	170	HP:000214	Frontotem
14	362	HP:000073	Disinhibitio
15	115	HP:000075	Personality
16	289	HP:000071	Agitation
17	108	HP:001196	Substantia
18	281	HP:000339	Muscle spa
19	589	HP:003022	Perseverat
20	308	HP:000235	Memory in
21	104	HP:000133	Dystonia

从泡泡图（两张泡泡图横向拼接成一张图进行展示！两张泡泡图和优化网络图是必贴的，最多贴三张图，还有一张看着贴就行。）中可以看出HPO基因集中与输入基因集关联最大的HPOterm就是帕金森，这证明了输入数据的质量还不错，进一步保证了富集结果的准确可靠；其次，还富集到了许多相关表型的terms，这有助于我们进一步理解帕金森的表型特征。

网络图（仅展示优化版）可以看出padj值最小的15个HPOterms与其他terms之间的联系，这有助于我们更详细地了解相关联的表型特征并进行扩展分析。

尽管通过HPO获得的表型结果分析是比较初步的，但是结果表格中不仅存在表型特征，还存在着各个terms中的关联基因，而这些基因可能会为对帕金森病的进一步研究做出贡献，如作为GO富集的输入基因集，进行进一步的关于分子层面的分析等等。

前往HPO官网查阅HP编号获取详细注释，padj值最小的前20个结果中的部分结果如下：

以下注释结果中存在一部分与帕金森高度重合的表型特征，主要反映了帕金森的症状。也有一些帕金森病的并发症表型特征被富集出来。

Parkinsonism HP:0001300

由中脑黑质多巴胺生成细胞变性引起的特征性神经异常，临床表现为震颤、僵直、运动迟缓、行走和步态困难。

Bradykinesia HP:0002067

运动迟缓(Bradykinesia)字面意思是运动缓慢，临床上用来表示运动执行缓慢(与运动减退(hypokinesia)相对，后者用于表示运动启动缓慢)。

Rigidity HP:0002063

连续不随意持续肌肉收缩。当受影响的肌肉被被动拉伸时，阻力的程度保持不变，与肌肉被拉伸的速率无关。这一特征有助于区分僵直和肌肉痉挛。

Postural instability HP:0002172

跌倒倾向或无力使自己不跌倒;不平衡。反推试验被广泛认为是评价姿势不稳定的金标准，反推试验的使用包括向后方向的快速平衡扰动，平衡纠正步骤的数量(或完全没有)被用来评价姿势不稳定的程度。健康的受试者可以通过一到两大步，或者不迈步来纠正这种干扰，他们可以快速地摆动手臂来平衡臀部。在平衡障碍患者中，平衡矫正步骤往往太小，迫使患者走两步以上。走3步或3步以上通常被认为是不正常的，走5步以上被认为是明显不正常的。明显受影响的患者继续后退，没有恢复平衡，必须由检查者抓住(这将被称为真正的后退)。更严重的患者不能完全纠正，像被推的玩具士兵一样向后摔倒，没有采取任何纠正措施。

Dementia HP:0000726

丧失整体认知能力的丧失，足以干扰正常的社会或职业功能痴呆通常是成年人原有认知能力的丧失，会影响记忆、思维、语言、判断和行为。

Resting tremor HP:0002322

当肌肉处于静止状态时，就会发生静止性震颤，当受影响的肌肉移动时，就会变得不那么明显或消失。静息性震颤通常是缓慢而粗糙的。静息性震颤会随着主动运动而消失，但通常会在手臂伸出几秒钟后再次出现(再次发生性震颤)。帕金森病休息震颤频率多为中低(3- 6hz)，振幅变化较大，宽度从小于1cm到超过10cm。

Depressivity HP:0000716

经常感到沮丧、痛苦和/或绝望;难以从这种情绪中恢复过来;对未来的悲观;普遍的耻辱;自卑的自我价值感;自杀的想法和自杀行为。

Parkinsonism with favorable response to dopaminergic medication HP:0002548

帕金森氏症是一种临床综合征，是许多不同疾病的特征，包括帕金森氏症本身，其他神经退行性疾病，如进行性核上麻痹，以及作为一些神经镇静药的副作用。一些但不是所有的帕金森病患者对多巴胺能药物有反应性，多巴胺能药物治疗后帕金森病的主要体征(主要包括震颤、运动迟缓、僵硬和姿势不稳)的改善显著减少。

Frontotemporal dementia HP:0002145

一种与额颞叶退化相关的痴呆，临床上与性格和行为变化如去抑制、冷漠和缺乏洞察力相关。额颞叶痴呆的显著特征是出现局灶性综合征，如进行性语言功能障碍、失语症或额叶功能障碍的行为改变。

Disinhibition HP:0000734

缺乏约束表现在几个方面，包括无视社会习俗、冲动和糟糕的风险评估。去抑制影响运动、本能、情绪、认知和知觉方面，其体征和症状类似于躁狂的诊断标准。性欲亢进，暴饮暴食，以及攻击性的爆发都是不受抑制的本能冲动的表现。

Personality changes HP:0000751

反常的转变思维、行动或感觉模式上的反常转变。这个术语指的是被认为是不正常的人格变化。它并不是指通常伴随年龄增长和某些生活状况而发生的性格变化。

Agitation HP:0000713

精神紧张与精神痛苦或内心紧张感相关的极度不安和过度运动的状态。

Substantia nigra gliosis HP:0011960

黑质胶质细胞局灶性增生。

Muscle spasm HP:0003394

肌肉的突然和不自觉的收缩一个或多个肌肉的突然和不自觉的收缩。

Perseveration HP:0030223

执拗可以被定义为上下文不恰当和无意识地重复一个反应或行为单元。换句话说，观察到的重复性不符合情境的要求，不是深思熟虑的产物，甚至可能在反意图的情况下展开。因此，执拗可以与目标导向的和有意的重复形式区分开来，比如旨在增强交际的语言。

Memory impairment HP:0002354

记忆力受损，表现为对日期和名字等事物的记忆能力下降，健忘增加。受影响的人倾向于在谈话中失去思路，开始任务但忘记他们的意图，在谈话中经常重复事情，并在日常生活任务中有困难。

Dystonia HP:0001332

肌张力异常增加导致固定的异常姿势有一个缓慢的，间歇性的扭转运动，导致夸大的转身和姿势的四肢和躯干。

后记（有需要的话自行自由补充）

课程论文构思以及攥写过程

5月中旬老师在微信群里发布HPO的辅助文件时本项目已经定下题目并且完成了基本的代码流程实现，因此我们优先选择了继续按照已有思路进行论文攥写。

我们的思路是先收集HPO富集分析的相关论文和辅助包，获取相关文献和帮助并且对比各个方法的可行性和难易度，最终决定了本次的分析方法。尽管目前我们也查阅到了更好的，依赖于R包的实验方法，但是其源代码过于复杂，且R的代码编写我们并不如Python熟悉，我们不能够在结课前完成对源代码的理解和分析，因此综合考虑后我们还是决定使用本文这种我们易于掌握的方法。

在实际实施中，遇到了一些代码问题和环境配置问题。得益于算法比较简单易于理解，代码问题在查阅源代码并进行增删修改后都得到了妥善解决。而环境问题主要是因为我一开始想用ubuntu20.0.4wsl（一个win10下的ubuntu子系统，能与win10同时运行）进行代码编译导致的，这个环境配置与一般的linux不太一样，而我又没有性能足够的linux服务器，所以花费了较多时间。然而在本论文中最终还是选择了win10环境执行代码。

参考资源或数据源网址

请见我的另一篇笔记：HPO富集分析项目笔记

我自己写的部分的GitHub网址仍在构建中，构建完成后会再加入此部分

生物信息学实验设计的构思与体会

帕金森是比较有名且常见的老年病，在决定采用这个方向的基因集作为输入数据时，其实预期的更多的是与其他没那么容易百度到的表型特征或者是与其他疾病的关联。但是从实际上的分析结果来看，排名前二十个的HPOterms给出的表型基本上都是可以直接在搜索引擎上搜索到的。解决这个问题的方法有两种，一个是继续查询p值更大的HPOterm注释，一个是从结果表格中筛选掉与帕金森深度绑定的terms。但是受限于篇幅和时间，我们可能会在结课后继续沿着这两个方向进行研究，继续深入理解帕金森的表型特征。

人员分工

实验设计，数据收集，代码设计，论文起草：生信1802杨晓龙

早期资料收集，正式论文攥写与排版：周旺