The instruction of enrichment calculation software for omics data

Abbreviation: oppOntology

Version 1.0

Designer: Zhang Yang

MENU

1. Introduction

Omics Pilot Platform of Ontology (oppOntology) is a kind of enrichment analysis tool for omics data with a graphical user interface (GUI), founded on the architecture of MATLAB AppDesigner. oppOntology supports the enrichment calculation of GO (Gene Ontology), KEGG (Kyoto Encyclopedia of Genes and Genomes), HPO (Human Phenotype Ontology), COG (Clusters of Orthologous Groups), MsigDB (Molecular Signatures Database) and Function (custom functions). Enrichment contraposes function categories of specific genes which contains the calculation of count, gene ratio, enrich factor, hypergeometry test, Fisher test, etc. This oppOntology supports the enrichment degree calculation of multiple samples simultaneously, and diagraming bubble graphs of all samples and various scores. The databases of GO, KEGG, COG and HPO are built-in databases in oppOntology whereas the data in database of MsigDB as the format of gmt could be downloaded in the official website of GSEA (http://www.gsea-msigdb.org/gsea/downloads.jsp). The files of Function comprised the text files of Function ID, Function Name and Gene ID, of which the users could define the text in the files to obtain the enrichment degree of the molecular function as their interest. oppOntology is straightforward for users and suitable for all the versions after MATLAB r2021. In the environment of MATLAB, the oppOntology could be started into the operation interface when the users type in the command “oppOntology”.

2. GUI of oppOntology

Open the MATLAB. In the panel of “APP”， click “install APP”. And then, select “oppOntology.mlappinstall” to load the package of oppOntology into MATLAB.

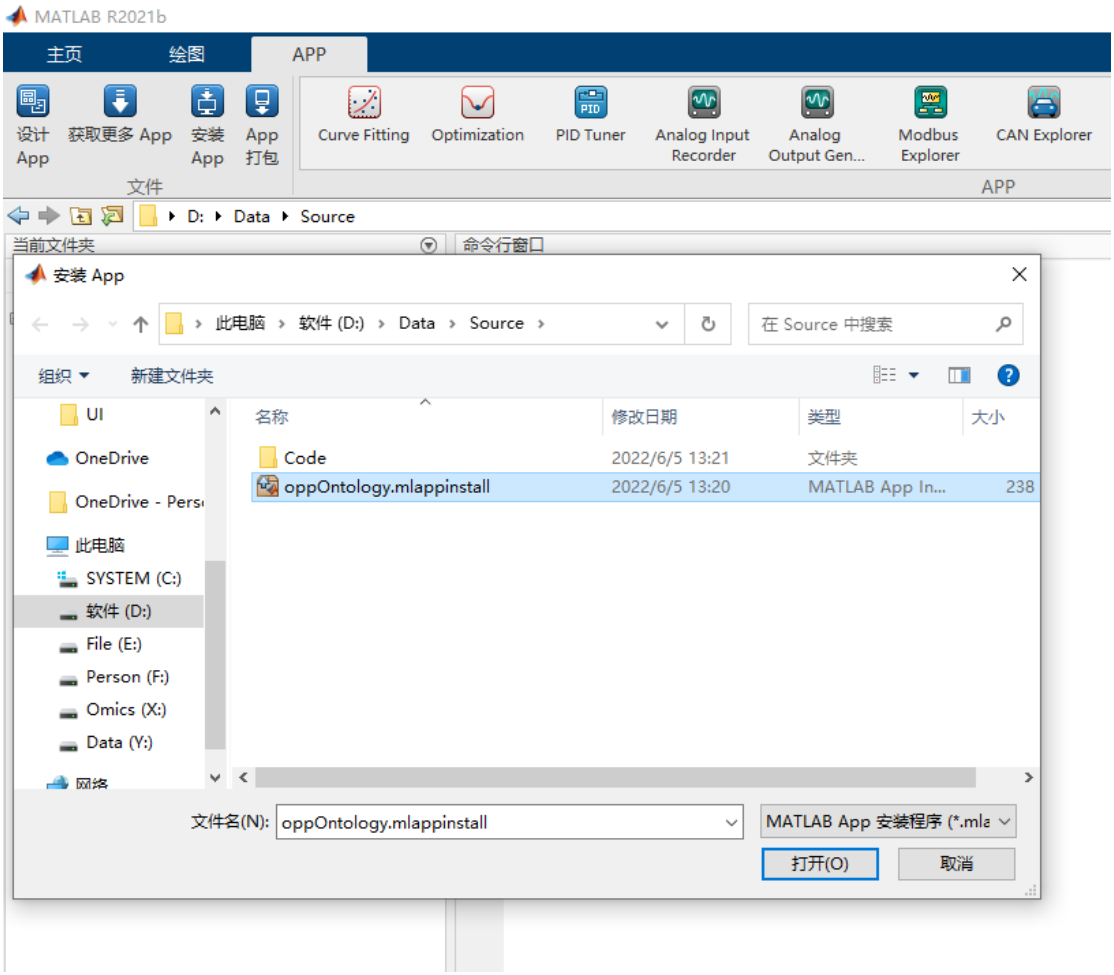

Figure 1. The installing procedure of oppOntology in MATLAB.

Execute the command “oppOntology” in the command region of MATLAB.

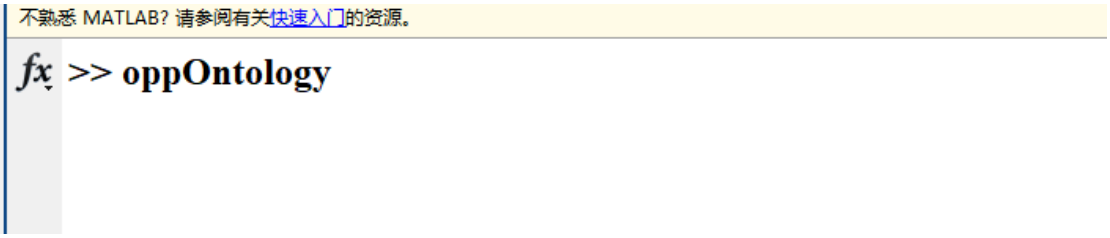

Figure 2. Start the program package of oppOntology.

Enter the main interface of oppOntology. There are 6 main interfaces containing Function, GO, KEGG, COG, HPO and MSigDB.

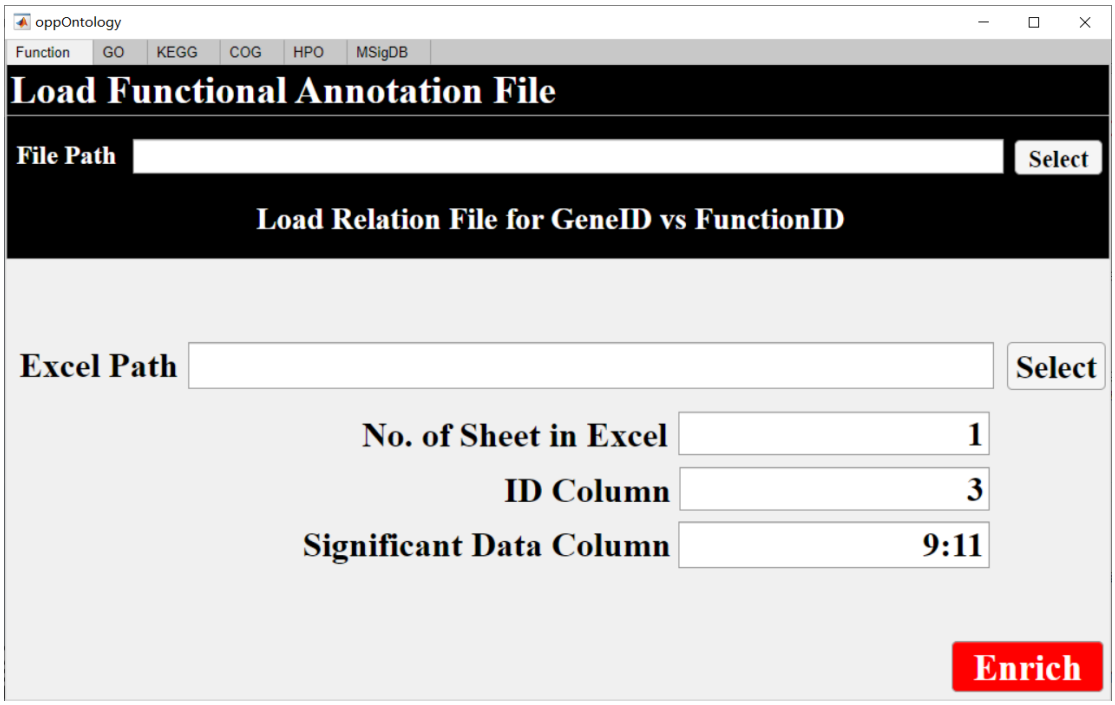

Figure 3. The main interface of oppOntology.

3. The Interface of Function

In the panel of Function, click “Select” to set the file path for analyzing the custom function in the format of txt. In the below panel of “Excel Path”, click “Select” to set the file path of original data in the format of Microsoft Excel, like “Data.xlsx”. In the panel of “No. of Sheet in Excel”, choose the candidate sheet according to the need of user, like typing number “1” represents the first sheet in the imported Excel file. In the panel of “ID Column”, choose the particular column as the gene IDs in the input data, like typing “3” represents the third column in the imported Excel file. In the panel of “Significant Data Column”, choose the particular columns for analyzing in the selected sheet. The type-in format is the numbers of first and last columns connected with a colon “:”.

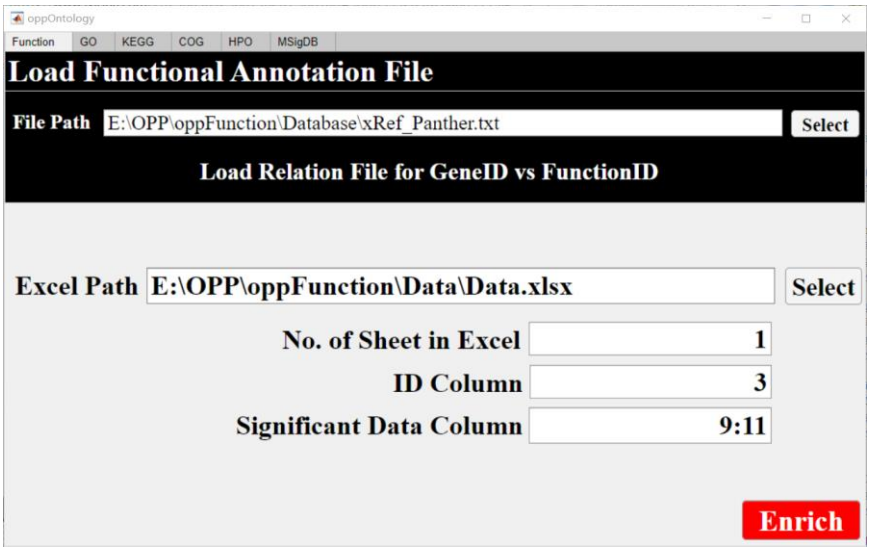

Figure 4. The parameter GUI of operating Function.

The files of Function should be comprised of Function IDs, Function Names, and Gene ID columns. The users could define the text in the files to obtain the enrichment degree of the molecular function as their interest.

FunctionID	FunctionName	ID
PTHR10000	PHOSPHOSERINE PHOSPHATASE	SERB_HUMAN
PTHR10003	SUPEROXIDE DISMUTASE [CU-ZN]-RELATED	CCS_HUMAN
PTHR10003	SUPEROXIDE DISMUTASE [CU-ZN]-RELATED	SODC_HUMAN
PTHR10003	SUPEROXIDE DISMUTASE [CU-ZN]-RELATED	SODE_HUMAN
PTHR10005	SKI ONCOGENE-RELATED	SKIL_HUMAN
PTHR10005	SKI ONCOGENE-RELATED	SKI_HUMAN
PTHR10005	SKI ONCOGENE-RELATED	SKOR1_HUMAN
PTHR10005	SKI ONCOGENE-RELATED	SKOR2_HUMAN
PTHR10006	MUCIN-1-RELATED	MUC1_HUMAN

Figure 5. The sample file for displaying essential file format of Function files.

The format of original data in the Microsoft Excel should contain the id numbers in the first few columns, the particular columns of storing significance data (Sig columns) and anything else in the middle columns. It should be mentioned that in the columns of significance data, these columns are made up of the numbers -1, 0, or 1. The number -1 represents downregulation, the number 1 represents upregulation, and the number 0 represents invariability.

1	2	3	4	5	6	7	8	9	10	11
GeneID	**Symbol**	**ID**	**Accession**	**ProteinName**	**LogRatio(2/1)**	**LogRatio(3/1)**	**LogRatio(4/1)**	**Sig(2/1)**	**Sig(3/1)**	**Sig(4/1)**
gene:23061	TBC1D9B	TBC9B_HUMAN	Q66K14	TBC1 domain famil	0.562546943	-0.110733597	-0.812557808	0	0	0
gene:29080	CCDC59	TAP26_HUMAN	Q9P031	Thyroid transcriptior	0.741506242	1.887766545	1.09045683	0	0	1
gene:29087	THYN1	THYN1_HUMAN	Q9P016	Thymocyte nuclear	-2.203859373	1.248095977	0.345259422	0	0	0
gene:79228	THOC6	THOC6_HUMAN	Q86W42	THO complex subu	-0.809063949	0.509922921	0.332358789	0	0	0
gene:7094	TLN1	TLN1_HUMAN	Q9Y490	Talin-1	-0.693475545	1.492975435	8.55E-02	0	1	0
gene:10693	CCT6B	TCPW_HUMAN	Q92526	T-complex protein 1	-0.42083631	1.311088189	0.626712219	0	1	0
gene:96764	TGS1	TGS1_HUMAN	Q96RS0	Trimethylguanosine	1.649804295	-2.177481916	1.750450427	0	0	0
gene:93643	TJAP1	TJAP1_HUMAN	Q5JTD0	Tight junction-assoc	-1.334556785	8.09E-02	0.142926379	0	0	0
gene:7023	TFAP4	TFAP4_HUMAN	Q01664	Transcription factor	1.399418475	2.357181294	0.631408509	0	0	0
gene:127262	TPRG1L	TPRGL_HUMAN	Q5T0D9	Tumor protein p63-r	-1.871528473	9.08E-02	-0.749305855	0	0	0
gene:79583	TMEM231	TM231_HUMAN	Q9H6L2	Transmembrane pro	-0.844902325	10	-0.494668556	0	0	0
gene:4234	METTL1	TRMB_HUMAN	Q9UBP6	tRNA (guanine-N(7	-1.443566432	2.626579863	0.25171689	-1	0	0
gene:122553	TRAPPC6B	TPC6B_HUMAN	Q86SZ2	Trafficking protein p	1.533285381	0.380803841	1.126153352	0	0	0
gene:23170	TTLL12	TTL12_HUMAN	Q14166	Tubulin--tyrosine liga	-0.324068984	0.766769073	-0.16695466	0	0	0
gene:7431	VIM	VIME_HUMAN	P08670	Vimentin	-0.374236872	-0.568591515	0.171903163	0	0	0
gene:65268	WNK2	WNK2_HUMAN	Q9Y3S1	Serine/threonine-pro	-1.055665608	2.414680744	-1.902586054	0	0	0
gene:65125	WNK1	WNK1_HUMAN	Q9H4A3	Serine/threonine-pro	-0.43446112	2.553242153	-0.300109473	0	0	0
gene:7520	XRCC5	XRCC5_HUMAN	P13010	X-ray repair cross-c	-0.424434772	1.145454224	9.60E-03	0	0	0
gene:738	VPS51	VPS51_HUMAN	Q9UID3	Vacuolar protein sor	0.665469272	1.151996856	0.281095267	0	0	0
gene:284403	WDR62	WDR62_HUMAN	O43379	WD repeat-containi	0.78628715	9.65E-02	2.299065019	0	0	0
gene:23230	VPS13A	VP13A_HUMAN	Q96RL7	Vacuolar protein sor	-2.583336219	-2.367745147	0.107850731	0	0	0
gene:23038	WDTC1	WDTC1_HUMAN	Q8N5D0	WD and tetratricope	-1.416677835	0.044643807	-1.048926407	0	0	0
gene:9203	ZMYM3	ZMYM3_HUMAN	Q14202	Zinc finger MYM-ty	0.149475453	-0.130175732	-1.483433162	0	0	0
gene:162963	ZNF610	ZN610_HUMAN	Q8N9Z0	Zinc finger protein 6	10	2.136136548	0.23646284	0	0	0
gene:51545	ZNF581	ZN581_HUMAN	Q9P0T4	Zinc finger protein 5	-3.096926518	0.026088476	0.410401286	0	0	0
gene:5859	QARS1	SYQ_HUMAN	P47897	Glutamine--tRNA lig	0.243374547	0.714691975	-4.99E-02	0	0	0
gene:81533	ITFG1	TIP_HUMAN	Q8TB96	T-cell immunomodul	-2.057185766	10	0.363623421	0	0	0
gene:80775	TMEM177	TM177_HUMAN	Q53SS8	Transmembrane pro	-1.795495316	-2.109307298	0.965444103	0	-1	0
gene:147007	TMEM199	TM199_HUMAN	Q8N511	Transmembrane pro	1.117566828	-2.888668917	0.584154955	0	0	0
gene:7050	TGIF1	TGIF1_HUMAN	Q15583	Homeobox protein T	-1.119668553	-1.86E-02	1.808684942	0	0	0

Figure 6. The sample file for displaying the format of original data in the Microsoft Excel.

Click the button “Enrich”, a new file folder named “Funciton” will be constructed in the current path. This file folder will include three files named “All”, “Down” and “Up”, and a file named Merge in the file format of “.xlsx”, “Merge.xlsx”. The file of “Merge.xlsx” will be the outcomes of the complete matrix correlation between original data and functions, which contains the original data in the first few columns and the following columns (named the function types) consist of the numbers in the extent of (0,1) as Boolean tables. The users could directly filter these raw data of their interest according to the built-in function of filtering in the Microsoft Excel. In the file folders of “All”, “Down” and “Up”, there will construct two new files “Samples.xlsx” and “Scores.xlsx” for storing data of enrichment degree, and two new figures according to data in these two files. Each sheet in the file of Samples.xlsx will provide enrichment degree calculated values of count, gene ratio, enrich factor, hypergeometry test, and Fisher test according to the function categories of specific genes listed in the **Sig** columns. Each sheet in the file of Scores.xlsx, there will provide the enrichment degree calculated scores of each sample listed in the **Sig** columns.

FID	Name	Count	GeneRatio	EnrichFactor	HyperTest	FisherTest
PTHR24073	FAMILY NOT NAMED	49	0.710144928	1.48760721	6.98125E-22	0.03789522
PTHR23239	INTERMEDIATE FILAMENT	38	0.542857143	1.137173791	1.8471E-06	0.536759382
PTHR24006	FAMILY NOT NAMED	35	0.593220339	1.242674304	4.10926E-07	0.320620321
PTHR24031	FAMILY NOT NAMED	35	0.875	1.832944598	2.73199E-16	0.009331933
PTHR24103	TRIM/RBCC (RING FINGER,	35	0.479452055	1.004353204	0.00092441	1
PTHR24115	FAMILY NOT NAMED	34	0.790697674	1.656348607	8.60934E-13	0.028367994
PTHR24089	FAMILY NOT NAMED	33	0.66	1.382563925	2.24737E-08	0.158463915
PTHR11915	SPECTRIN/FILAMIN RELATI	32	0.8	1.675835061	5.31935E-12	0.031755045
PTHR10799	SWI/SNF-RELATED MATRIX	31	0.911764706	1.909959077	6.01889E-15	0.011063888
PTHR13140	MYOSIN	30	0.625	1.309246141	1.96215E-06	0.27492627
PTHR22957	TBC1 DOMAIN FAMILY MEI	30	0.681818182	1.428268518	7.38051E-08	0.136048687
PTHR24361	MITOGEN-ACTIVATED KIN	30	0.810810811	1.698481481	2.76865E-11	0.035716114
PTHR24377	FAMILY NOT NAMED	29	0.202797203	0.424818328	0	5.67332E-06
PTHR24412	FAMILY NOT NAMED	28	0.528301887	1.106683531	0.000891012	0.721240983
PTHR14047	FAMILY NOT NAMED	26	0.684210526	1.433279986	1.42169E-06	0.180207502
PTHR24012	FAMILY NOT NAMED	26	0.702702703	1.472017283	5.74418E-07	0.138301823
PTHR24416	TYROSINE-PROTEIN KINAS	26	0.481481481	1.008604435	0.012277825	1
PTHR11254	HECT DOMAIN UBIQUITIN-	22	0.846153846	1.772517853	2.19996E-08	0.062270897
PTHR19134	PROTEIN-TYROSINE PHOSF	22	0.611111111	1.280151783	0.000348848	0.399092509
PTHR24067	UBIQUITIN-CONJUGATING	22	0.75862069	1.589153937	1.03281E-06	0.101123059

Figure 7. The sample file for displaying the output file format of Samples.xlsx.

FID	Name	Sample	Sig(2/1)	Sig(3/1)	Sig(4/1)
PTHR24073	FAMILY NOT N	0.03789522	0.052887722	0.110448596	0.013903446
PTHR23239	INTERMEDIAT	0.536759382	0.543133781	0.69660855	0.468168033
PTHR24006	FAMILY NOT N	0.320620321	0.370004393	0.255815501	0.359838342
PTHR24031	FAMILY NOT N	0.009331933	0.001347755	0.059525223	0.00044158
PTHR24103	TRIM/RBCC (R	1	0.84526369	0.922004326	1
PTHR24115	FAMILY NOT N	0.028367994	0.00649862	0.003600998	0.013703556
PTHR24089	FAMILY NOT N	0.158463915	0.328072631	0.545867017	0.208114999
PTHR11915	SPECTRIN/FILA	0.031755045	0.00727337	0.007373514	0.114855446
PTHR10799	SWI/SNF-RELA	0.011063888	0.000116368	0.000119435	0.000965149
PTHR13140	MYOSIN	0.27492627	0.126050266	0.455880381	0.198714444
PTHR22957	TBC1 DOMAIN	0.136048687	0.079684411	0.055937904	0.097667376
PTHR24361	MITOGEN-ACT	0.035716114	0.051821984	0.004274495	0.029730576
PTHR24377	FAMILY NOT N	5.67332E-06	0.010998397	0.005250908	0.001990557
PTHR24412	FAMILY NOT N	0.721240983	0.643566765	0.559901477	0.724826508
PTHR14047	FAMILY NOT N	0.180207502	0.483565843	0.07741308	0.105185022
PTHR24012	FAMILY NOT N	0.138301823	0.193010032	0.101428106	0.135195472
PTHR24416	TYROSINE-PR(	1	0.910203607	1	1
PTHR11254	HECT DOMAIN	0.062270897	0.042329082	0.007339529	0.011944068
PTHR19134	PROTEIN-TYR(	0.399092509	0.471494194	0.67413246	0.373459278
PTHR24067	UBIQUITIN-CC	0.101123059	0.417559277	0.323554458	0.243496699

Figure 8. The sample file for displaying the output file format of Scores.xlsx.

Figure 9. The simple figure for displaying the function of diagraming bubble graph according to the data in Sig columns of Samples.xlsx and Scores.xlsx.

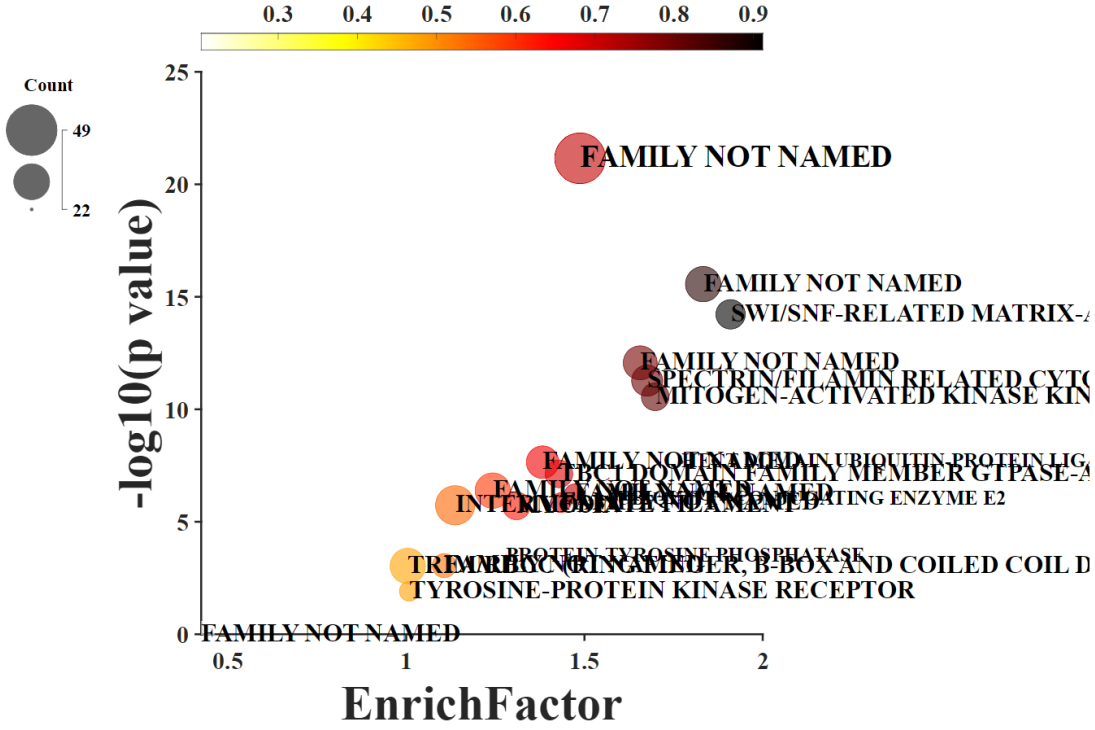

Figure 10. The simple figure for displaying the function of demonstrating the several enrichment degrees values according to the data in specific Sig columns of Samples.xlsx and Scores.xlsx.

4. Function Analysis of GO

4.1 Annotation

In the panel of GO's Annotation, click "Select" to set the file path of obo annotated files for GO. The obo annotated files are the files that contain all the GO annotated numbers provided by GO official website. In the below list box of "GO ID List", users could type in the numbers of GO IDs. The button "Get" in the panel of "Get Annotation" is used to obtain the annotation names of input GO IDs including Field, Name and Description. The button "Get" in the panel of "Get Ancestors" is used to obtain the ancestor GO IDs of input GO IDs. The button "Get" in the panel of "Get Descendants" is used to obtain the descendant GO IDs of input GO IDs. The button "Get" in the panel of "Get Relatives" is used to obtain the relative GO IDs of input GO IDs. All the outcomes will be generated in a file of GO.xlsx in the current path.

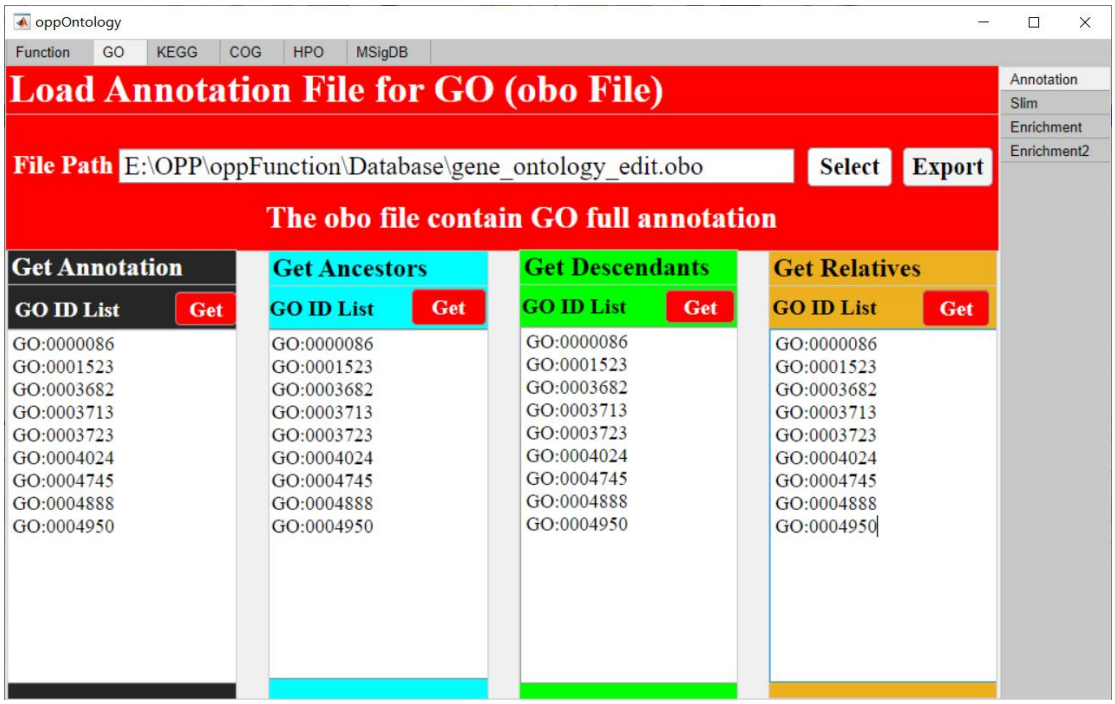

Figure 11. The parameter GUI of operating GO's Annotation.

4.2 Enrichment

In the panel of GO's Enrichment, click "Select" to set the file path of gene numbers associated with GO to upload GO cross-reference file, like xRef_GO.txt. In the below panel of "Excel Path", click "Select" to set the file path of original data in file format of Microsoft Excel, like "Data.xlsx". In the panel of "No. of Sheet in Excel", choose the candidate sheet according to the need of user, like typing number "1" represents the first sheet in the imported Excel file. In the panel of "ID Column", choose the particular column as the gene IDs in the input data, like typing "3" represents the third column in the imported Excel file. In the panel of "Significant Data Column", choose the particular columns for analyzing in the selected sheet. The type-in format is the numbers of first and last columns connected with a colon ":". Click the button "Enrich",

a new file folder named “Function” will be constructed in the current path. This file folder will include three files named “All”, “Down” and “Up”, and a file named Merge in the file format of “.xlsx”, “Merge.xlsx”. The interpretation of results is same as the meaning of former “Function”, which need not repeat it here.

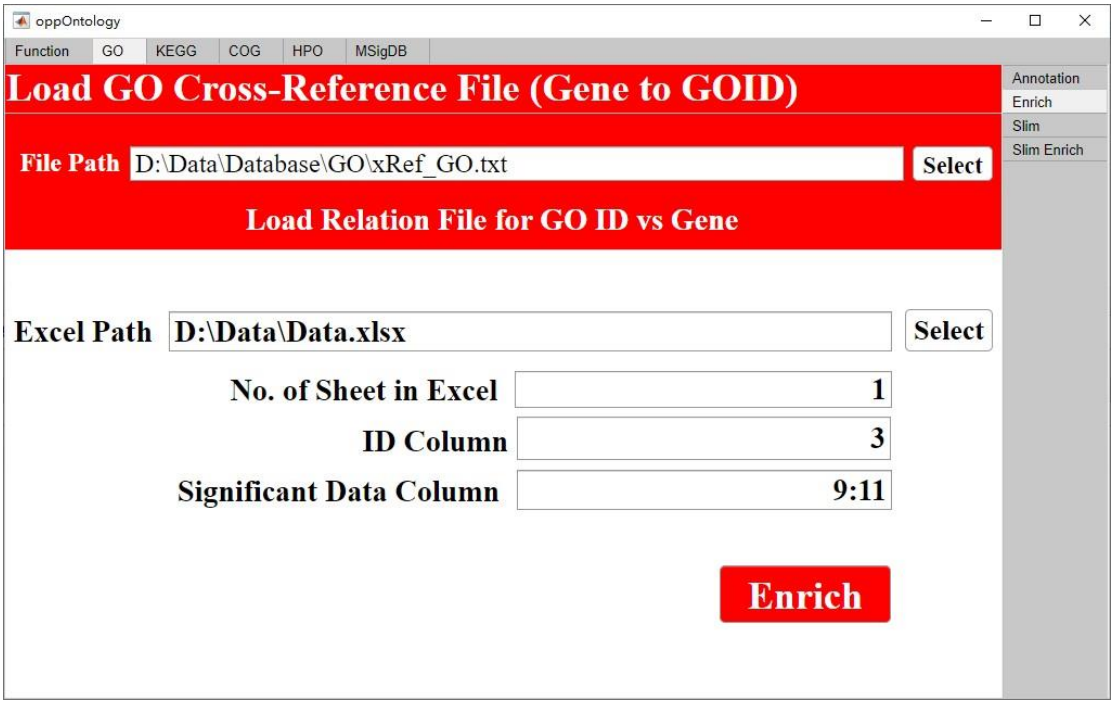

Figure 12. The parameter GUI of operating GO’s Enrichment.

4.3 Slim

Under the panel of GO’s Slim, in the panel of “Load Annotation File for GO (obo File)”, click “Select” to set the file path of obo annotated files for GO. The obo annotated files are the files that contain all the GO annotated numbers provided by GO official website. In the middle panel of “GO Cross-Reference File (Gene to GOID)”, click “Select” to set the file path of gene numbers associated with GO to upload GO cross-reference file, like xRef_GO.txt. In the bottom panel of “Load Parents for GO (Excel File)”, click “Select” to set the file path of user-defined parents IDs for GO, like “Parent.xlsx”.

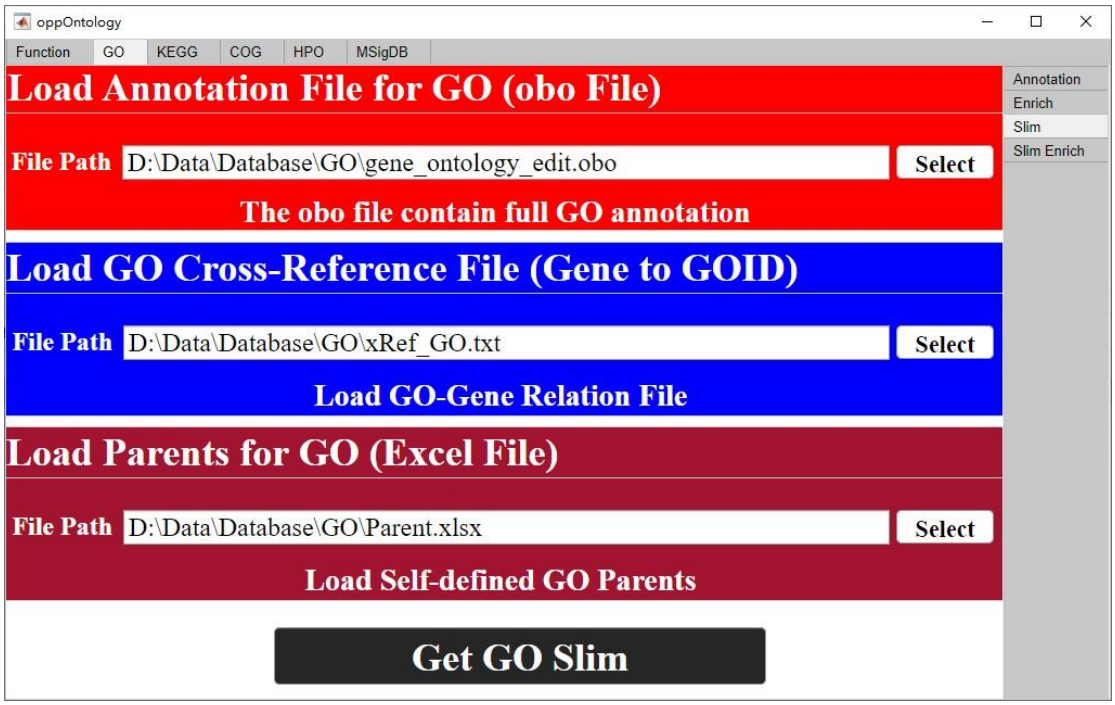

Figure 13. The parameter GUI of operating GO's Slim.

The file format of "xRef_GO.txt" was shown in the following figure. The columns of ID, GOID, Field and Name are essential.

ID	GOID	Field	Name	Evidence
ACKR4_HUMAN	GO:0005623	C	cell	IEA:GOC
ACKR4_HUMAN	GO:0005769	C	early endosome	IEA:UniProtKB-SubCell
ACKR4_HUMAN	GO:0009897	C	external side of plasma membrane	IBA:GO_Central
ACKR4_HUMAN	GO:0005887	C	integral component of plasma membrane	TAS:ProtInc
ACKR4_HUMAN	GO:0005886	C	plasma membrane	TAS:Reactome
ACKR4_HUMAN	GO:0055037	C	recycling endosome	IEA:UniProtKB-SubCell
ACKR4_HUMAN	GO:0019957	F	C-C chemokine binding	IBA:GO_Central
ACKR4_HUMAN	GO:0016493	F	C-C chemokine receptor activity	IBA:GO_Central
ACKR4_HUMAN	GO:0019956	F	chemokine binding	IBA:GO_Central
ACKR4_HUMAN	GO:0004950	F	chemokine receptor activity	IDA:UniProtKB
ACKR4_HUMAN	GO:0005044	F	scavenger receptor activity	IEA:InterPro
ACKR4_HUMAN	GO:0019722	P	calcium-mediated signaling	IBA:GO_Central
ACKR4_HUMAN	GO:0060326	P	cell chemotaxis	IBA:GO_Central

Figure 14. The sample figure for the file format of "xRef_GO.txt".

The files of "Parent.xlsx" should contain three sheets of CC, BP and MF. The format of these three sheets should be consistent. The sheet format of "CC" was shown in the following figure.

Field	SlimID	SlimName
C	GO:0005886	Plasma Membrane
C	GO:0016020	Membrane
C	GO:0005576	Extracellular Region
C	GO:0005634	Nucleus
C	GO:0005730	Nucleolus
C	GO:0005737	Cytoplasm
C	GO:0005739	Mitochondrion
C	GO:0005783	ER
C	GO:0005794	Golgi Apparatus
C	GO:0005764	Lysosome
C	GO:0005768	Endosome
C	GO:0005777	Peroxisome
C	GO:0005840	Ribosome
C	GO:0005811	Lipid Particle

Figure 15. The sample figure for the file format of "CC".

In the current path, a file named "SlimTable.xlsx" would be established. In the sheet of Slim, the correlation of GO IDs and Slim IDs would be involved.

ID	Field	GOID	SlimID	SlimName
ACKR4_HUMAN	C	GO:0005769	GO:0005737	Cytoplasm
ACKR4_HUMAN	C	GO:0005769	GO:0005768	Endosome
ACKR4_HUMAN	C	GO:0009897	GO:0005886	Plasma Membrane
ACKR4_HUMAN	C	GO:0009897	GO:0016020	Membrane
ACKR4_HUMAN	C	GO:0005887	GO:0005886	Plasma Membrane
ACKR4_HUMAN	C	GO:0005887	GO:0016020	Membrane
ACKR4_HUMAN	C	GO:0005886	GO:0005886	Plasma Membrane
ACKR4_HUMAN	C	GO:0005886	GO:0016020	Membrane
ACKR4_HUMAN	C	GO:0055037	GO:0005737	Cytoplasm
ACKR4_HUMAN	C	GO:0055037	GO:0005768	Endosome
3HIDH_HUMAN	C	GO:0005759	GO:0005737	Cytoplasm
3HIDH_HUMAN	C	GO:0005759	GO:0005739	Mitochondrion
3HIDH_HUMAN	C	GO:0005739	GO:0005737	Cytoplasm
3HIDH_HUMAN	C	GO:0005739	GO:0005739	Mitochondrion
1433G_HUMAN	C	GO:0005829	GO:0005737	Cytoplasm
1433G_HUMAN	C	GO:0070062	GO:0005576	Extracellular Region
1433G_HUMAN	C	GO:0016020	GO:0016020	Membrane
1433G_HUMAN	C	GO:0005739	GO:0005737	Cytoplasm
1433G_HUMAN	C	GO:0005739	GO:0005739	Mitochondrion
ACL6B_HUMAN	C	GO:0071565	GO:0005634	Nucleus

Figure 16. The sample figure for the file format of "SlimTable.xlsx".

4.4 Slim Enrich（Enrichment for SlimID）

Under the panel of GO's Slim Enrich, in the panel of "Load Go Slim Annotation Excel", click the button "Select" to set the file path of correlation of GO IDs and Slim IDs, like the file "SlimTable.xlsx" established in the Chapter 4.3. In the below panel of "Excel Path", click "Select" to set the file path of original data in file format of Microsoft Excel, like "Data.xlsx". In the panel of "No. of Sheet in Excel", choose the candidate sheet according to the need of user, like typing number "1" represents the first sheet in the imported Excel file. In the panel of "ID Column", choose the particular column as the gene IDs in the input data, like typing "3" represents the third column in the imported Excel file. In the panel of "Significant Data Column", choose the particular columns for analyzing in the selected sheet. The type-in format is the numbers of first and last columns connected with a colon ":".

Click the button "Enrich", a new file folder named "GO Slim" will be constructed in the current path. This file folder will include three files named "All", "Down" and "Up", and a file named Merge in the file format of ".xlsx", "Merge.xlsx". The interpretation of results is same as the meaning of former "Function", which need not repeat it here.

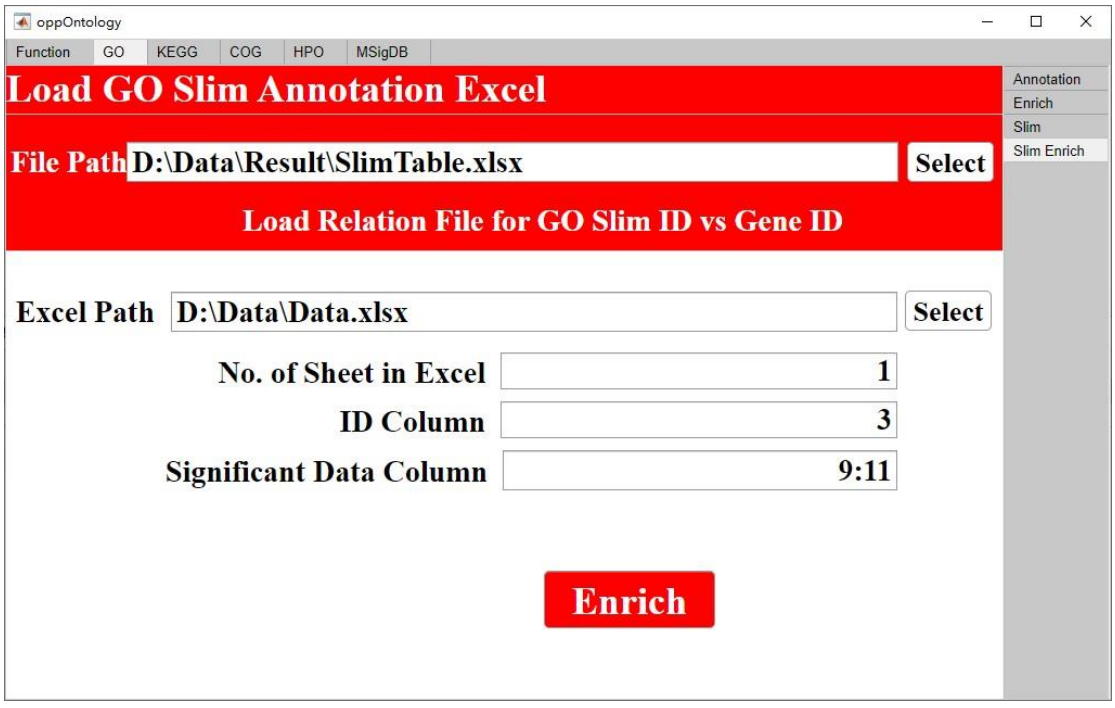

Figure 17. The parameter GUI of operating GO's Slim Enrich.

5. KEGG pathway analysis

In the panel of KEGG, click the button "Select" to set the file path for correlation of KEGG and gene IDs, like xRef_KO.txt. In the below panel of "Excel Path", click "Select" to set the file path of original data in file format of Microsoft Excel, like "Data.xlsx". In the panel of "No. of Sheet in Excel", choose the candidate sheet

according to the need of user, like typing number “1” represents the first sheet in the imported Excel file. In the panel of “ID Column”, choose the particular column as the gene IDs in the input data, like typing “3” represents the third column in the imported Excel file. In the panel of “Significant Data Column”, choose the particular columns for analyzing in the selected sheet. The type-in format is the numbers of first and last columns connected with a colon “:”.

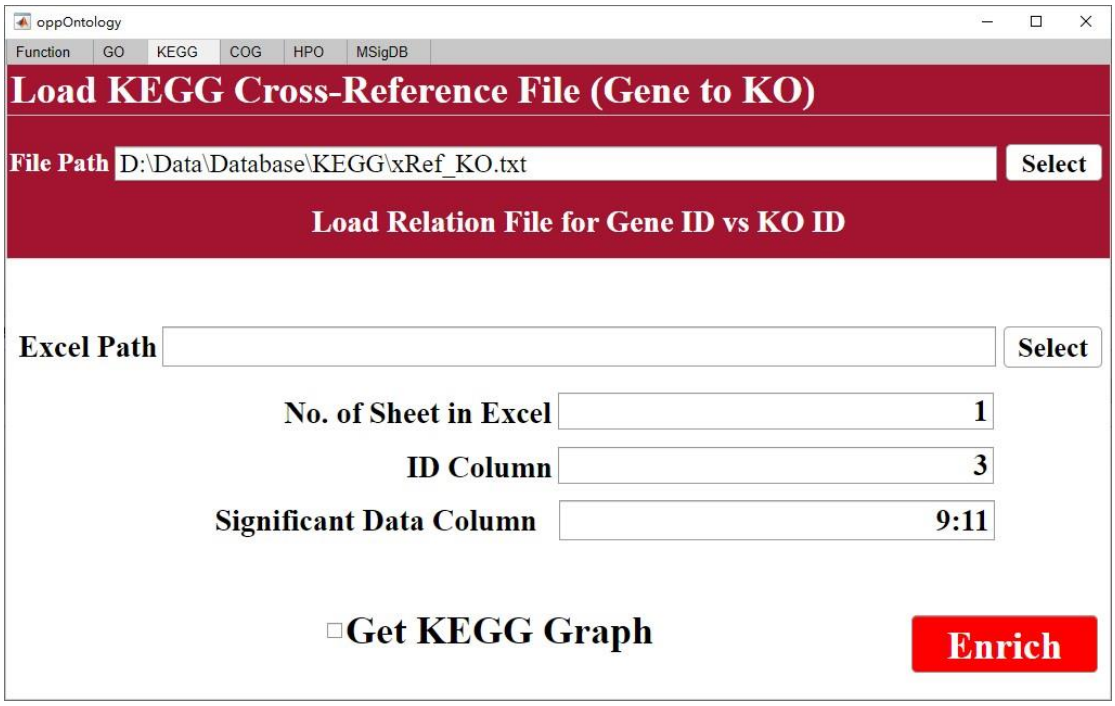

Figure 18. The parameter GUI of operating KEGG.

Click the button “Enrich”, a new file folder named “KEGG” will be constructed in the current path. This file folder will include three files named “All”, “Down” and “Up”, and a file named Merge in the file format of “.xlsx”, “Merge.xlsx”. The interpretation of results is same as the meaning of former “Function”, which need not repeat it here. Moreover, a file folder named “KEGG\Graph” will also established, which, according to the column of Significant Data Column, contains the figures mapping to pathways in the website of KEGG. In the KEGG pathway figures, red colors represent upregulation genes in the column of Significant Data Column while green colors represent downregulation genes.

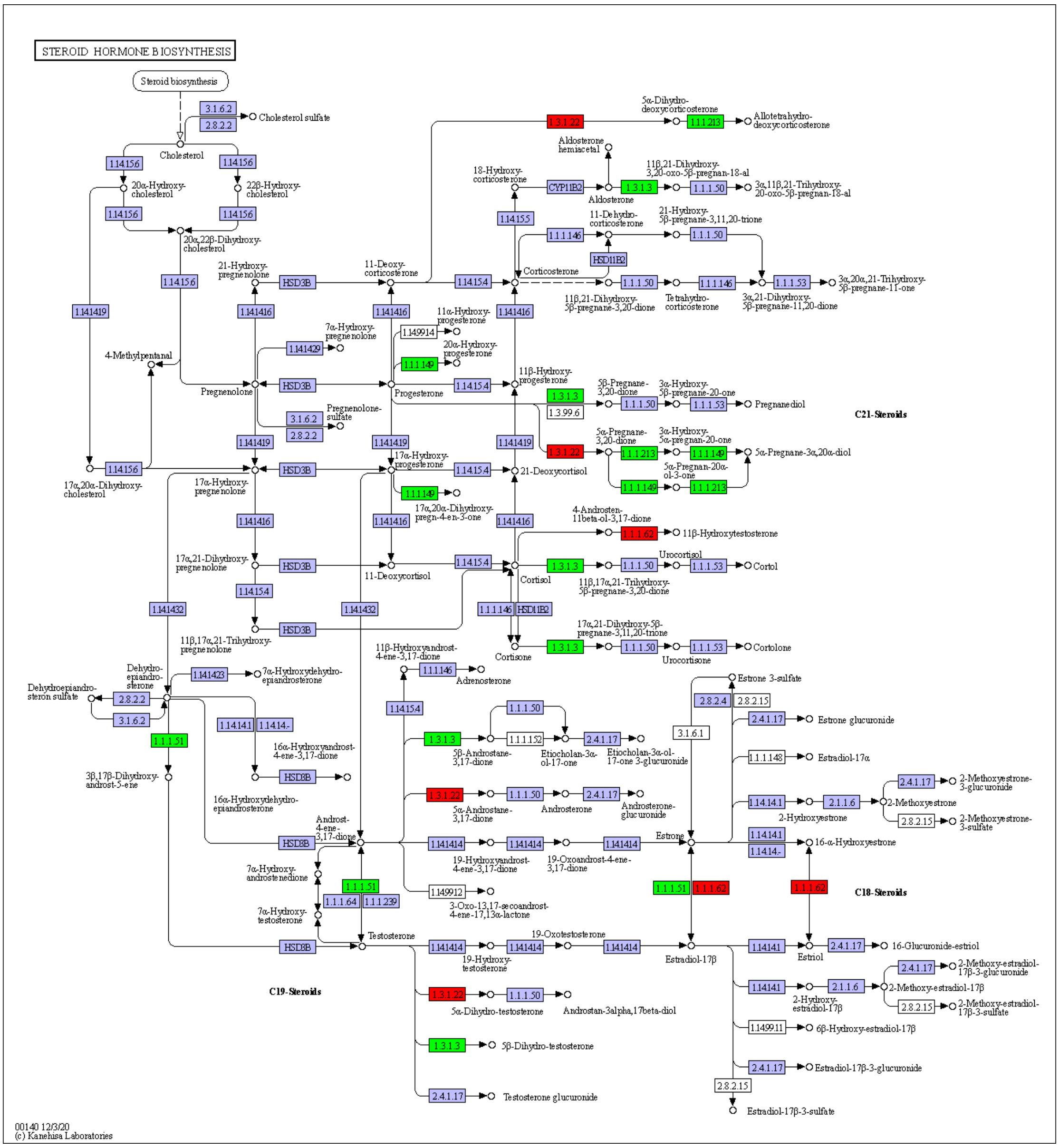

Figure 19.　The sample figure of outcomes in KEGG pathway analysis.

6. Function Analysis of COG

In the panel of COG, we provided the function to analyze correlation and summary statistics of gene lists through COG (Clusters of Orthologous Groups of proteins). In the panel of “Load COG Annotation Excel File”, click “Select” to set the file path of COG databases information which contains different levels of COG annotation. The information of COG should be saved in the sheet of “ID” of upload file, like “COG.xlsx”. As the built-in function in MATLAB is to call the BLAST program online, which is with a slow speed, the BLAST+ matched results are also acceptable in this software. In the panel of “Load BLAST + Search Result (txt format, -outfmt 6)”, click “Select” to set the file path to upload the BLAST+ matched results.

The reference commands for BLAST+ are as follows:

>Blastp -outfmt 6 -db COG Path/COG.fasta -query Query Path/Protein.fasta -out BlastOut.txt

Herein, the value of the parameter -outfmt is 6, and the file format. of output is tsv.

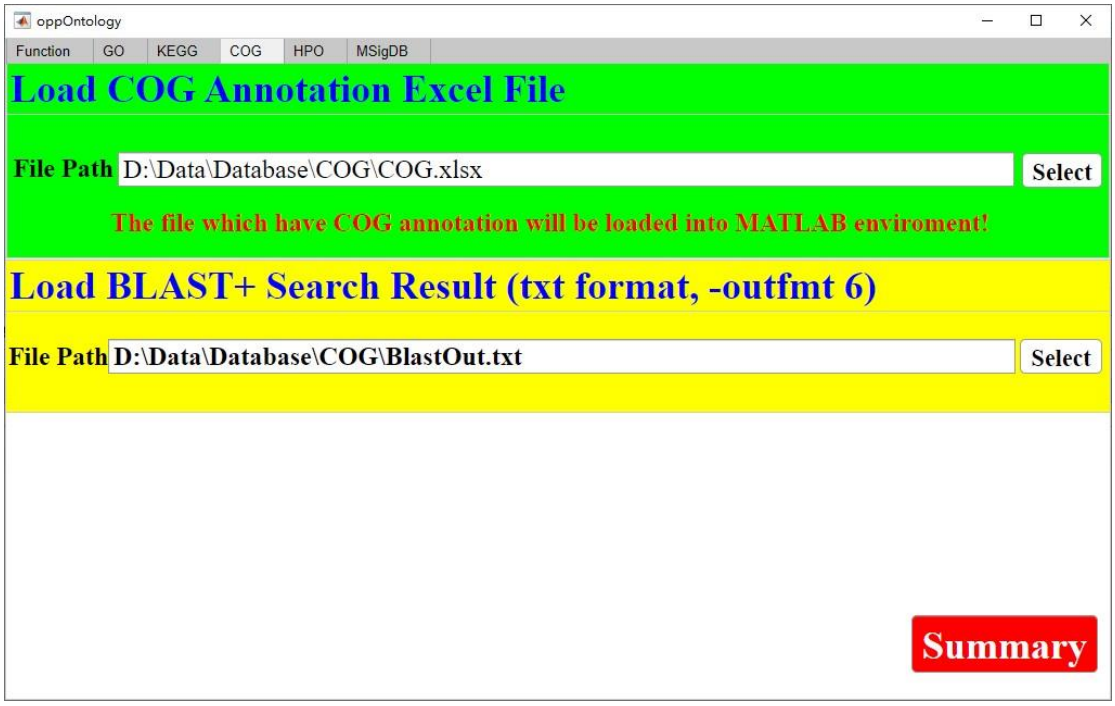

Figure 20. The parameter GUI of operating COG.

First, the outcomes of BALST+ will be matched to the COG databases to form a matched table.

Class	Symbol	Name	Subject	Query	Score	eValue
CELLULAR PROCESSES AND SIGNALI	M	Cell wall/membrane/envelope biogenesis	CAC3503	TR100922|c0_g1|m.4246	50.8	1.64E-06
CELLULAR PROCESSES AND SIGNALI	M	Cell wall/membrane/envelope biogenesis	SPBC3F6.02c	TR101039|c0_g2|m.22834	197	8.91E-58
CELLULAR PROCESSES AND SIGNALI	M	Cell wall/membrane/envelope biogenesis	sll0045_1	TR100991|c0_g1|m.21354	327	2.7E-105
CELLULAR PROCESSES AND SIGNALI	M	Cell wall/membrane/envelope biogenesis	sll0045_1	TR100991|c0_g4|m.27916	65.1	4.55E-13
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	BH0831	TR101036|c0_g6|m.3402	58.5	1.41E-09
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	BS_vpr	TR101036|c0_g16|m.43780	64.3	8.14E-13
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	BS_vpr	TR101036|c0_g4|m.48496	62	3.42E-11
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	BS_vpr	TR101036|c0_g5|m.25660	44.7	7.28E-06
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	CC1107	TR100020|c0_g1|m.63201	255	1.88E-78
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	CC2505	TR100998|c0_g1|m.37429	117	3.17E-34
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	ECU11g1130	TR100926|c0_g1|m.39267	50.4	3.17E-07
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	NMB1027_1	TR101024|c0_g5|m.20204	70.9	1.6E-14
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	PM1024	TR100987|c2_g1|m.39283	187	1.49E-54
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	RSc0384	TR100353|c0_g4|m.34876	106	2.72E-29
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	SPAC13G7.02c	TR101043|c2_g12|m.29456	74.7	1.1E-16
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	SPAC13G7.02c	TR101043|c2_g19|m.45233	135	5.9E-38
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	SPAC1527.03	TR100018|c0_g1|m.54460	68.2	7.27E-12
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	SPAC1527.03	TR100972|c0_g3|m.19340	81.3	2.8E-15
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	SPAC323.02c	TR100285|c0_g5|m.52780	295	1.1E-101
CELLULAR PROCESSES AND SIGNALI	O	Posttranslational modification, protein turnover, chaperones	SPAC323.02c	TR100317|c0_g6|m.30609	338	3.8E-118

Figure 21. The sample figure of BALST+ matched table.

Then, summary statistics will be operated, and graphs will be plotted in different levels of COG.

Symbol	Name	GroupCount
A	RNA processing and modification	2
C	Energy production and conversion	12
E	Amino acid transport and metabolism	22
G	Carbohydrate transport and metabolism	20
H	Coenzyme transport and metabolism	11
I	Lipid transport and metabolism	15
J	Translation, ribosomal structure and biogenesis	29
K	Transcription	27
L	Replication, recombination and repair	19
M	Cell wall/membrane/envelope biogenesis	4
O	Posttranslational modification, protein turnover, chaperones	34
P	Inorganic ion transport and metabolism	14
Q	Secondary metabolites biosynthesis, transport and catabolism	5
R	General function prediction only	33
S	Function unknown	4
T	Signal transduction mechanisms	17
U	Intracellular trafficking, secretion, and vesicular transport	2
Z	Cytoskeleton	6

Figure 22. The sample figure of summary statistics.

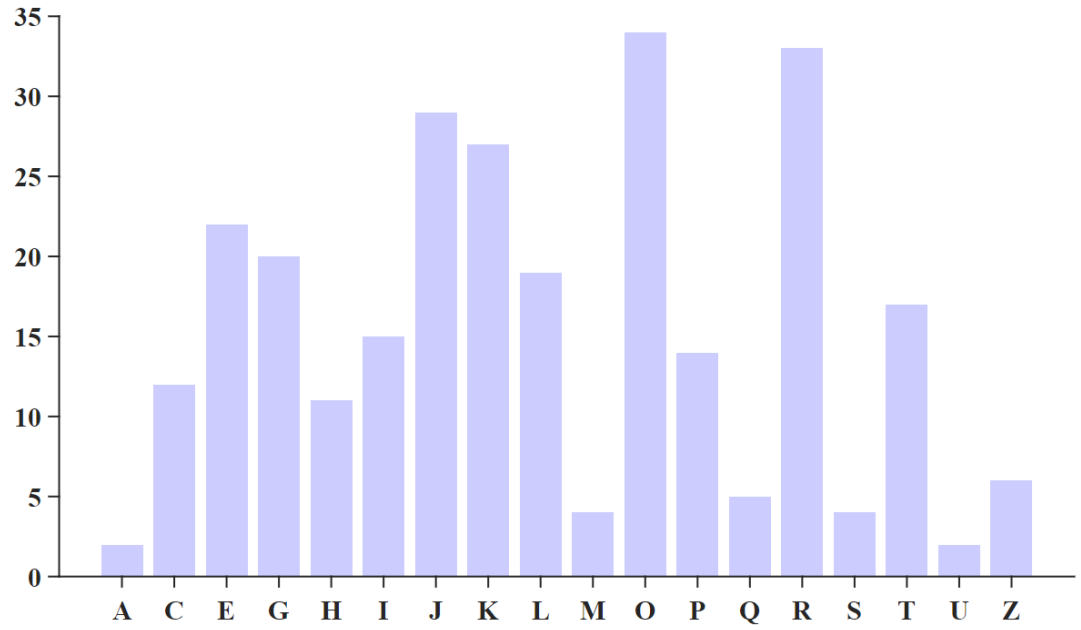

Figure 23. The sample figure of plotted graph.

7. Function Analysis of HPO

In the panel of HPO, click the button “Select” to set the file path for correlation of HPO and gene IDs, like xRef_KO.txt. In the below panel of “Excel Path”, click “Select”

to set the file path of original data in file format of Microsoft Excel, like “Data.xlsx”. In the panel of “No. of Sheet in Excel”, choose the candidate sheet according to the need of user, like typing number “1” represents the first sheet in the imported Excel file. In the panel of “ID Column”, choose the particular column as the gene IDs in the input data, like typing “3” represents the third column in the imported Excel file. In the panel of “Significant Data Column”, choose the particular columns for analyzing in the selected sheet. The type-in format is the numbers of first and last columns connected with a colon “:”.

Click the button “Enrich”, a new file folder named “HPO” will be constructed in the current path. This file folder will include three files named “All”, “Down” and “Up”, and a file named Merge in the file format of “.xlsx”, “Merge.xlsx”. The interpretation of results is same as the meaning of former “Function”, which need not repeat it here.

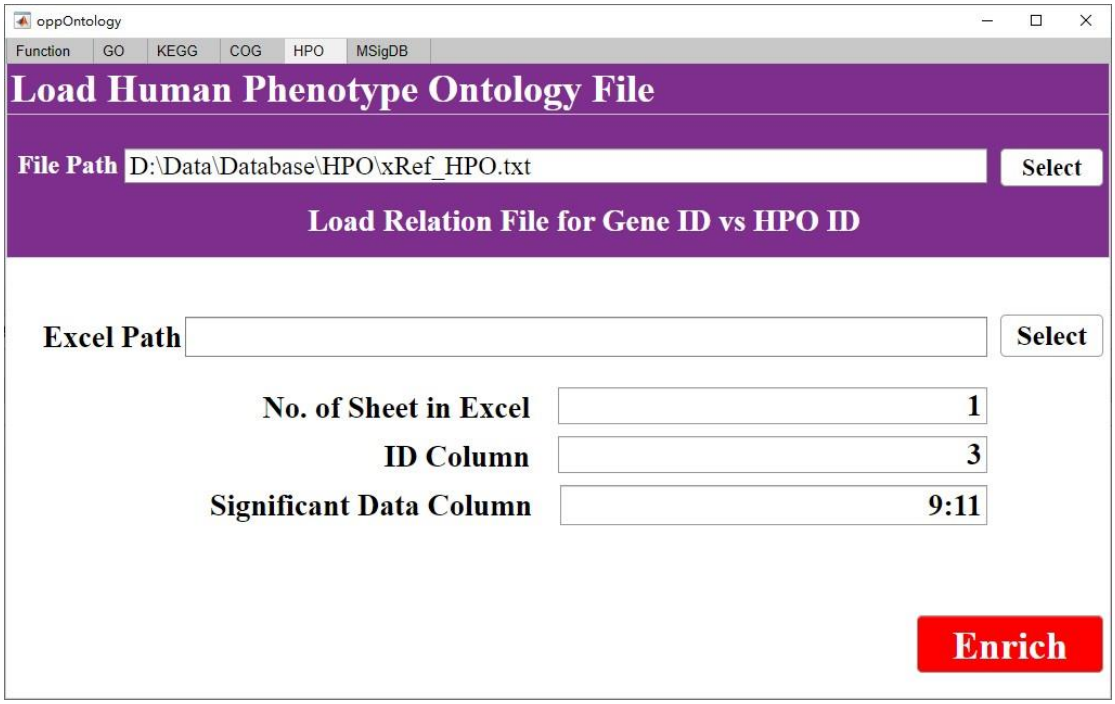

Figure 24. The parameter GUI of operating HPO.

8. Function Analysis of MSigDB

In the panel of MSigDB, click the button “Select” to set the file path for correlation of MSigDB and gene IDs, like *.gmt. *.gmt could be downloaded in the official website of GSEA (https://www.gsea-msigdb.org/gsea/index.jsp). Click the button “Export” to convert the file format of gmt into table format. In the below panel of “Excel Path”, click “Select” to set the file path of original data in file format of Microsoft Excel, like “Data.xlsx”. In the panel of “No. of Sheet in Excel”, choose the candidate sheet according to the need of user, like typing number “1” represents the first sheet in the imported Excel file. In the panel of “ID Column”, choose the particular column as the

gene IDs in the input data, like typing “3” represents the third column in the imported Excel file. In the panel of “Significant Data Column”, choose the particular columns for analyzing in the selected sheet. The type-in format is the numbers of first and last columns connected with a colon “:”.

Click the button “Enrich”, a new file folder named “MSigDB” will be constructed in the current path. This file folder will include three files named “All”, “Down” and “Up”, and a file named Merge in the file format of “.xlsx”, “Merge.xlsx”. The interpretation of results is same as the meaning of former “Function”, which need not repeat it here.

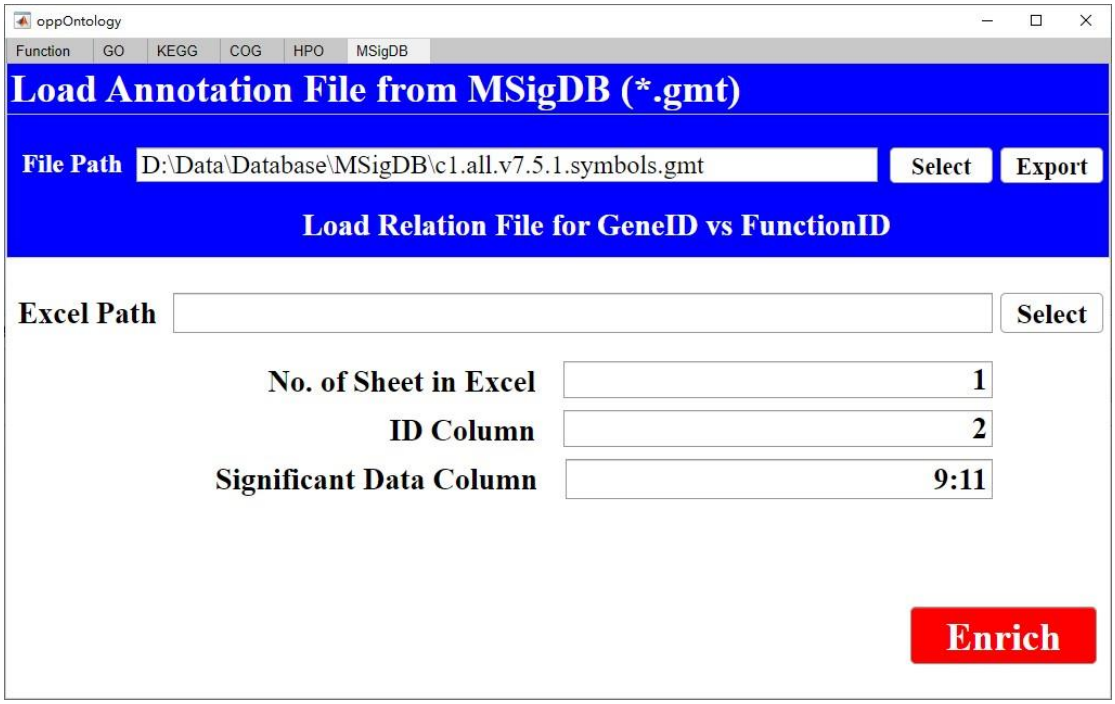

Figure 25. The parameter GUI of operating MSigDB.