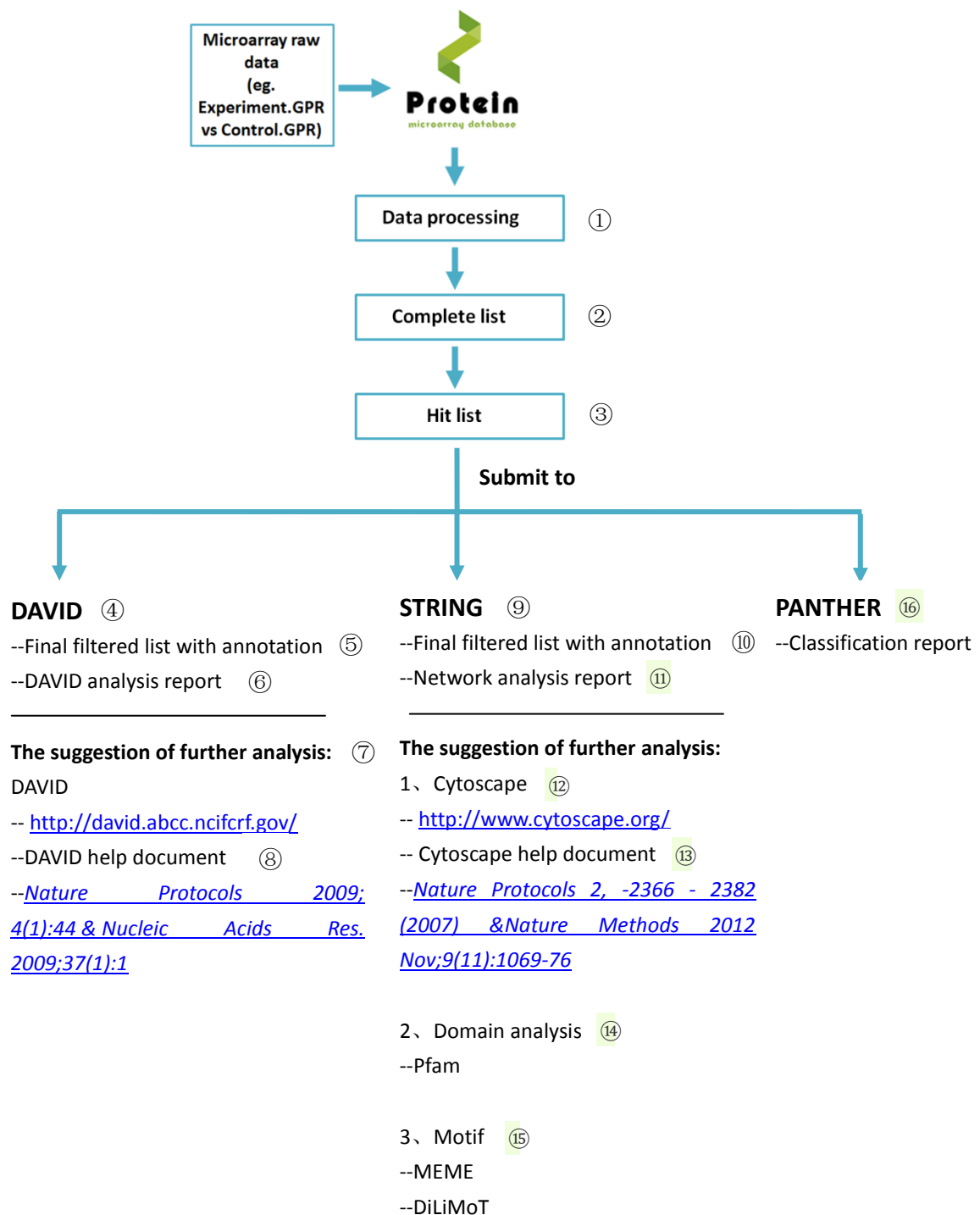


How PMD performs protein microarray data analysis

Protein Microarray Database (PMD) also provides tools for protein microarray data processing and basic bioinformatical analysis. You can simply upload paired microarray raw data, eg. Experiment.GPR vs Control.GPR (or select from what you have up loaded) and click a button, then enjoy!

You can learn the analysis process in the flowchart, and detailed notes are shown as follow.



Besides there processing above, you also can try Docking ⑰ , eg. Protein vs Small molecular, Protein vs Protein, and so on.

Notes:

① Data processing

After reading microarray raw data, the R (1)page of limma(2) and MASS(3) are applied to standardize this raw data. The code is as follows.

```
backgroundCorrect(RG,method="normexp")
normalizeBetweenArrays(RG.bgcorrect)
```

② Complete list

In complete list, the information of protein list, standardized data, fold change and p value is included. (see fig.1) The standardized data is calculated for fold change between experiment and control. In addition Z-test is used to discriminate the difference. The code is as follows.

```
fold change = log2(x/y)
p <- 1 - pnorm(x, y)
```

If the microarray data number is greater than 2, fisher test will be used. The code is as follows.

```
fisher.test(m, alternative = "less")$p.value
fisher.test(m, alternative = "greater")$p.value
```

1		Experiment	Control	fold change	p value
2	A2M	4.275393	5.1034625	-0.255419	0.7961844
3	A2ML1	4.7007564	4.9312619	-0.069064	0.5911505
4	A4GALT	4.7485582	4.8528059	-0.03133	0.5415136
5	AAA1	4.8919603	4.7599802	0.0394571	0.4475
6	AAAS	4.3366212	4.365664	-0.00963	0.5115648
7	AACS	6.4779437	4.51638	0.520369	0.0249067
8	AADAC	4.7228067	4.9810645	-0.07681	0.601896
9	AADACL2	4.9332641	4.3729348	0.1739467	0.2676274
10	AADAT	4.2621563	4.7756074	-0.164101	0.6961821
11	AAK1	4.7617024	4.6722121	0.0273717	0.4643461
12	AAMP	4.7357018	4.744353	-0.002633	0.5034513
13	AARSD1	4.656688	4.6383135	0.0057039	0.4926701
14	AASDH	3.2135088	4.8284231	-0.587403	0.9468354
15	AASDHPTT	4.7633983	4.6689104	0.0289053	0.4623608
16	AATF	4.2229529	6.1535684	-0.543171	0.9732347
17	AATK	4.9964078	4.6689104	0.0978053	0.3716459
18	ABAT	4.4233651	4.8062043	-0.119754	0.6490805
19	ABCA10	3.9202223	4.3706983	-0.156928	0.6738164
20	ABCA12	6.5487631	4.9721395	0.3973557	0.0574411
21	ABCA2	4.7274892	4.6849663	0.0130355	0.4830409
22	ABCA5	4.7232883	4.7089748	0.0043786	0.4942899
23	ABCA6	4.0362195	5.9091633	-0.549949	0.9694619
24	ABCA8	4.9388598	4.613231	0.0984005	0.3723526
25	ABCA9	4.837219	5.3305624	-0.14011	0.689115
26	ABCB5	4.2931504	4.7137561	-0.13484	0.6629785
27	ABCB6	4.779165	4.8616086	-0.024675	0.532853
28	ABCB7	5.120391	4.3112571	0.2481454	0.2092191
29	ABCB9	5.0791047	4.726072	0.1039326	0.362032
30	ABCC1	4.8149087	4.9480605	-0.039355	0.5529633
31	ABCC10	4.9609781	5.1342206	-0.049521	0.5687696
32	ABCC12	5.0879755	5.4337388	-0.094854	0.6352397
33	ABCC5	4.0134728	4.8027596	-0.0060947	0.4917372

Z-test's p value

Fold change

Standardized data

Fig.1. The example of complete list.

③ Hit list

Differential protein, defined as p value less than or equal to 0.05 will be selected to hit list (see fig.2), and participate to next analysis. The code is as follows.

```
result[which(result[,c("p value")]<=0.05),]
```


⑤ DAVID final filtered list with annotation

When analyze in DAVID, a few of proteins are unmapped and the mapped proteins are show in DAVID final filtered list. The file includes Uniprot ID, protein name and species. (see fig.4)

1	UNIPROT_AC Name	Species	
2	P36896	activin A Homo sapiens	
3	Q15Q57	ArfGAP with Homo sapiens	
4	Q8WWZ4	ATP-bindir Homo sapiens	
5	O15254	acyl-CoenzHomo sapiens	
6	Q13685	angio-asscHomo sapiens	
7	Q9UG63	ATP-bindirHomo sapiens	
8	P08910	abhydrolasHomo sapiens	
9	Q9NUQ8	ATP-bindirHomo sapiens	
10	P49748	acyl-CoenzHomo sapiens	
11	P16442	ABO blood Homo sapiens	
12	Q9BTE6	alanyl-tRFHomo sapiens	
13	P00813	adenosine Homo sapiens	
14	Q8NC06	acyl-CoenzHomo sapiens	
15	Q99965	ADAM metalHomo sapiens	
16	O00590	chemokine Homo sapiens	
17	Q13444	ADAM metalHomo sapiens	
18	O75077	ADAM metalHomo sapiens	
19	Q8TC27	ADAM metalHomo sapiens	
20	P35318	adrenomedvHomo sapiens	
21	Q4L235	aminoadipsHomo sapiens	
22	Q9BZ11	ADAM metalHomo sapiens	
23	Q86V21	acetoacetyHomo sapiens	
24	P11766	alcohol deHomo sapiens	
25	P61158	ARF3 actirHomo sapiens	
26	Q8NOZ2	actin-bincHomo sapiens	
27	Q9NFI3	acyl-CoA tHomo sapiens	
28	O94805	actin-likeHomo sapiens	
29	O14678	ATP-bindirHomo sapiens	
30	Q8NFI4	abhydrolasHomo sapiens	
31	P40394	alcohol deHomo sapiens	
32	Q9UKF2	ADAM metalHomo sapiens	
33	Q9NFI5	acyl-CoA tHomo sapiens	

Uniprot ID

Protein name

Species

Fig.4. The final filtered list about DAVID.

⑥ DAVID analysis report

In order to demonstrate a neat format, we suggest that you open this analysis report with Microsoft Excel. You can get different information through the different display modes. Firstly, the annotation cluster can be getting in data frame. (see fig.5A) Secondly, different types of annotation will show in data frame though sorting the first column in ascending order. (see fig.5B) In order to make clear about DAVID result, PMD provide a visualized form. (see fig.5C)

A

The first column

1	Annotator Enrichment Score: 2.686040094678326				
2	Category	Term	Count	%	PValue
3	INTERPRO	IPRO13087:	21	1.124197	3.84E-06
4	UP_SEQ_FEA	zinc finger	19	1.0171306	6.69E-06
5	UP_SEQ_FEA	zinc finger	18	0.9635974	1.39E-05
6	UP_SEQ_FEA	zinc finger	17	0.9100642	1.60E-05
7	UP_SEQ_FEA	zinc finger	19	1.0171306	2.49E-05
8	INTERPRO	IPRO07087:	22	1.1777302	3.53E-05
9	INTERPRO	IPRO15880:	22	1.1777302	4.48E-05
10	SP_PIR_KEY	transcript	38	2.0342612	5.10E-05
11	SMART	SM00355:Zr	22	1.1777302	5.80E-05
12	SP_PIR_KEY	Transcript	38	2.0342612	8.10E-05
13	GOTERM_BP	GO:0006350	38	2.0342612	1.70E-04
14	UP_SEQ_FEA	zinc finger	16	0.856531	1.86E-04
15	UP_SEQ_FEA	zinc finger	17	0.9100642	2.13E-04
16					
17	Annotator Enrichment Score: 1.2464941955562054				
18	Category	Term	Count	%	PValue
19	GOTERM_BP	GO:0021546	3	0.1605996	0.0343985
20	GOTERM_BP	GO:0022037	3	0.1605996	0.0455965
21	GOTERM_BP	GO:0030902	3	0.1605996	0.1161574
22					
23	Annotator Enrichment Score: 1.2099498498765444				
24	Category	Term	Count	%	PValue
25	SP_PIR_KEY	synapse	7	0.3747323	0.0154266
26	SP_PIR_KEY	cell junct	9	0.4817987	0.0345657
27	GOTERM_CC	GO:0045202	7	0.3747323	0.0916853
28	GOTERM_CC	GO:0030054	7	0.3747323	0.3016585
29					
30	Annotator Enrichment Score: 0.8261518133098353				
31	Category	Term	Count	%	PValue
32	SP_PIR_KEY	repressor	9	0.4817987	0.0525097
33	GOTERM_MF	GO:0043565	10	0.5353319	0.1599696

Annotation Cluster 1

Annotation Cluster 2

Annotation Cluster 3

Annotation Cluster 4

B

109	Category	Term	Count	%	PValue
110	GOTERM_BP	GO:0006350	38	2.0342612	1.70E-04
111	GOTERM_BP	GO:0021546	3	0.1605996	0.0343985
112	GOTERM_BP	GO:0022037	3	0.1605996	0.0455965
113	GOTERM_BP	GO:0030902	3	0.1605996	0.1161574
114	GOTERM_BP	GO:0007411	5	0.267666	0.0206824
115	GOTERM_BP	GO:0030900	5	0.267666	0.0615681
116	GOTERM_BP	GO:0031175	6	0.3211991	0.1056296
117	GOTERM_BP	GO:0007406	5	0.267666	0.1201342
118	GOTERM_BP	GO:0001764	3	0.1605996	0.1290985
119	GOTERM_CC	GO:0045202	7	0.3747323	0.0916853
120	GOTERM_CC	GO:0030054	7	0.3747323	0.3016585
121	GOTERM_CC	GO:0019717	4	0.2141328	0.0585393
122	GOTERM_CC	GO:0005624	10	0.5353319	0.2779266
123	GOTERM_CC	GO:0005626	10	0.5353319	0.3142271
124	GOTERM_CC	GO:0000267	11	0.5888651	0.4798625
125	GOTERM_CC	GO:0005736	17	0.9100642	0.0272801
126	GOTERM_CC	GO:0005576	25	1.3383298	0.0665323
127	GOTERM_CC	GO:0030054	7	0.3747323	0.3016585
128	GOTERM_CC	GO:0005912	3	0.1605996	0.3849348
129	GOTERM_CC	GO:0070161	3	0.1605996	0.4464318
130	GOTERM_CC	GO:0005882	4	0.2141328	0.2163272
131	GOTERM_MF	GO:0043565	10	0.5353319	0.1599696
132	GOTERM_MF	GO:0018564	5	0.267666	0.3956106
133	GOTERM_MF	GO:0003700	16	0.856531	0.0664495
134	GOTERM_MF	GO:0003690	4	0.2141328	0.074803
135	GOTERM_MF	GO:0030526	21	1.124197	0.1244645
136	GOTERM_MF	GO:0043566	4	0.2141328	0.1804935
137	GOTERM_MF	GO:0018563	6	0.3211991	0.3984178
138	GOTERM_MF	GO:0005509	10	0.5353319	0.588503
139	GOTERM_MF	GO:0042803	6	0.3211991	0.2481149
140	GOTERM_MF	GO:0046983	8	0.4282855	0.3067911
141	GOTERM_MF	GO:0042806	6	0.3211991	0.2185577

GOTERM_BP_FAT

GOTERM_CC_FAT

GOTERM_MF_FAT

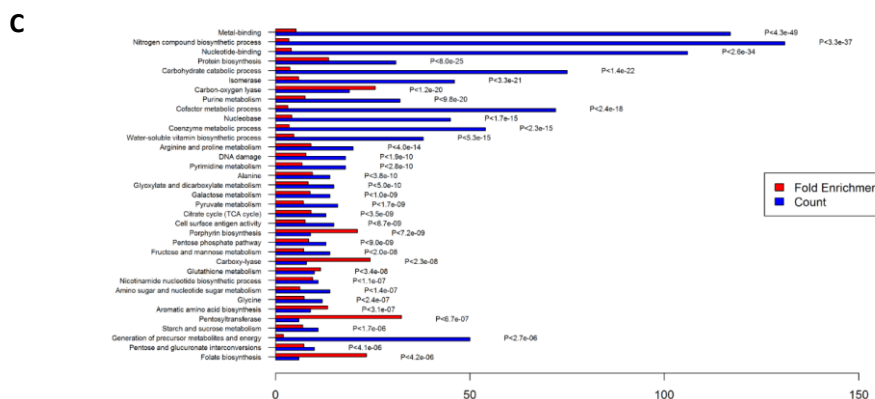


Fig.5. The example of DAVID analysis report. A. The annotation cluster. B. Different types of annotation. C. Visualizing the result of DAVID enrichment by histogram.

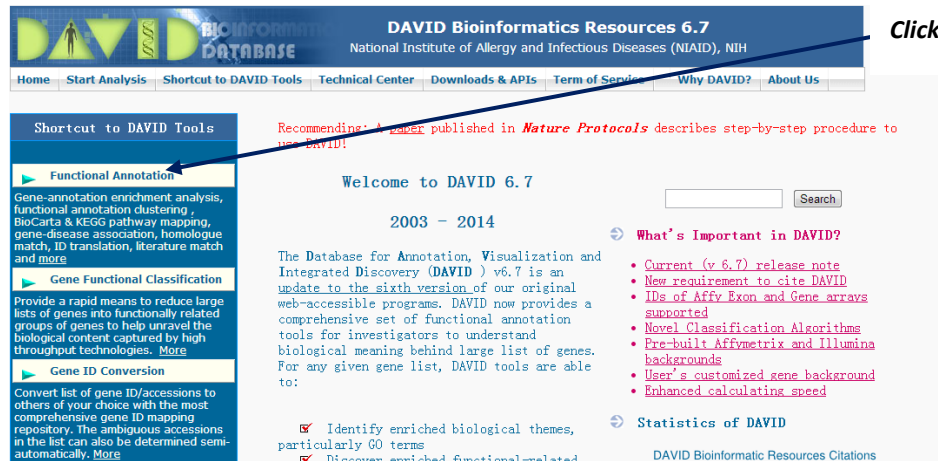
⑦ Further analysis

In further analysis, we provide some advice for helping users mining data.

⑧ DAVID help document

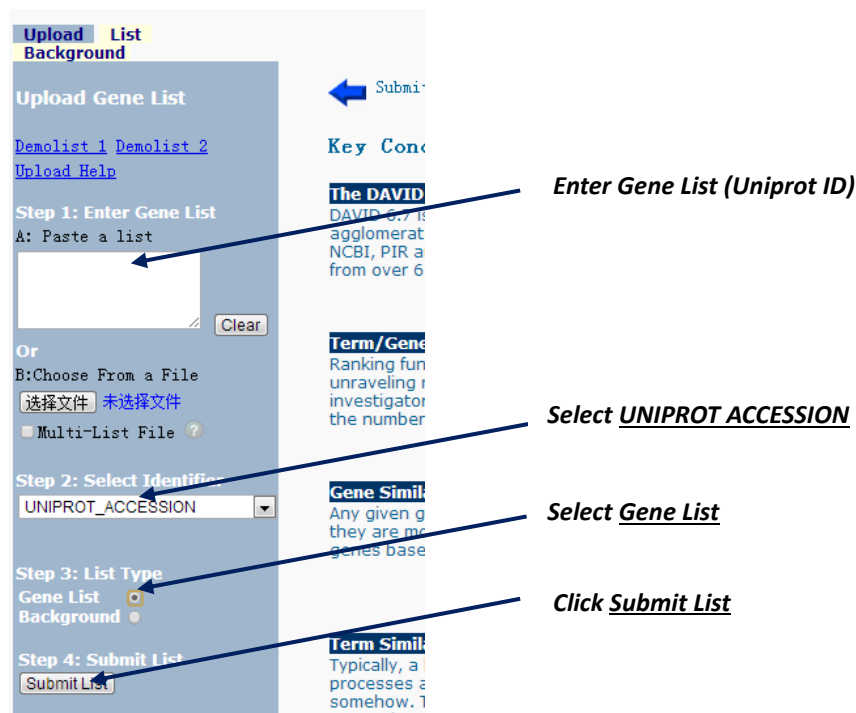
In this analysis process, the threshold is default value, so you can try re-analysis in DAVID by yourself, and the steps are as follows.

A、Open the URL: <http://david.abcc.ncifcrf.gov/>



Click Functional Annotation

B、



Enter Gene List (Uniprot ID)

Select UNIPROT ACCESSION

Select Gene List

Click Submit List

C、

DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Clustering

Current Gene List: List_1
Current Background: Homo sapiens
175 DAVID IDs

[Help Manual](#)

Options **Classification Stringency** Medium

Kappa Similarity **Similarity Term Overlap** 3 **Similarity Threshold** 0.5C

Classification **Initial Group Membership** 3 **Final Group Membership** 3 **Multiple Linkage Threshold** 0.5C

Enrichment Thresholds **EASE** 1.0

Display ☐ Fold Change ☐ Bonferroni ☒ Benjamini ☐ FDR ☐ LT,PH,PT

[Rerun using options](#) [Create Sublist](#)

63 Cluster(s) [Download File](#)

Annotation Cluster 1	Enrichment Score: 25.42	Count	P_Value	Benjamini
<input type="checkbox"/> INTERPRO	ABC transporter-like	28	1.6E-42	3.8E-40
<input type="checkbox"/> INTERPRO	ABC transporter, conserved site	28	9.1E-40	1.1E-37
<input type="checkbox"/> GOTERM_MF_FAT	purine nucleoside binding	76	1.0E-30	3.3E-28
<input type="checkbox"/> GOTERM_MF_FAT	nucleoside binding	76	1.6E-30	2.6E-28
<input type="checkbox"/> GOTERM_MF_FAT	adenyl nucleotide binding	75	2.9E-30	3.2E-28
<input type="checkbox"/> SP_PIR_KEYWORDS	ato-binding	61	7.1E-27	1.6E-24
<input type="checkbox"/> GOTERM_MF_FAT	purine nucleotide binding	75	8.7E-25	7.1E-23

Gene list being analyzed

Clustering options and stringency

The overall enrichment score for the group based on the EASE scores of each term members. The higher, the more enriched.

Every term in the annotation cluster

Related Term Search

Genes involved in individual term

ALL genes involved in this annotation cluster

A group of terms having similar biological meaning due to sharing similar gene members

Functional Annotation Clustering

Current Gene List: demolist1
171 DAVID IDs

[Rerun using options](#) [Create Sublist](#) [Download File](#)

Options **Classification Stringency** High

Annotation Cluster 1 **Enrichment Score: 3.69**

<input type="checkbox"/> SP_PIR_KEYWORDS	chromoprotein	RT	7	1.1E-5
<input type="checkbox"/> SP_PIR_KEYWORDS	metalloprotein	RT	8	4.7E-5
<input type="checkbox"/> SP_PIR_KEYWORDS	iron	RT	9	1.1E-4
<input type="checkbox"/> GOTERM_MF_ALL	iron ion binding	RT	10	2.5E-4
<input type="checkbox"/> SP_PIR_KEYWORDS	heme	RT	7	3.5E-4
<input type="checkbox"/> GOTERM_MF_ALL	tetrapyrrole binding	RT	6	1.3E-3
<input type="checkbox"/> GOTERM_MF_ALL	heme binding	RT	6	1.3E-3

Annotation Cluster 2 **Enrichment Score: 3.52**

<input type="checkbox"/> SP_PIR_KEYWORDS	antibiotic	RT	5	2.2E-4
<input type="checkbox"/> SP_PIR_KEYWORDS	antimicrobial	RT	5	2.4E-4
<input type="checkbox"/> GOTERM_BP_ALL	defense response to bacteria	RT	6	5.4E-4

Annotation Cluster 3 **Enrichment Score: 2.46**

<input type="checkbox"/> UP_SEQ_FEATURE	domain:Ig-like C2-type 1	RT	8	5.4E-4
<input type="checkbox"/> UP_SEQ_FEATURE	domain:Ig-like C2-type 2	RT	8	5.4E-4
<input type="checkbox"/> INTERPRO_NAME	immunoglobulin	RT	6	3.6E-2

Annotation Cluster 4 **Enrichment Score: 2.63**

<input type="checkbox"/> SP_PIR_KEYWORDS	immunoglobulin	RT	6	3.6E-2
--	----------------	----	---	--------

EASE Score, the modified Fisher Exact P-Value. They are identical to that in the Chart Report. The smaller, the more enriched.

Fig.6. DAVID help document.

⑨ STRING

PMD provides an analytic process that can output Protein-Protein interaction network (PPI-network). This process is set up in STRING that is connecting with PMD by a back-end Python relational program. STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations.

⑩ STRING final filtered list with annotation

When analyze in STRING, a few of proteins are unmapped and the mapped proteins are show

in STRING final filtered list. The file includes gene name, gene annotation. (see fig.7)

1	MT1E	9606. ENSPCmetallothi
2	FOXN1	9606. ENSPCforkhead b
3	ARPP19	9606. ENSPCcAMP-regul
4	ZNHIT1	9606. ENSPCzinc finge
5	ZSCAN1	9606. ENSPCzinc finge
6	ZNF193	9606. ENSPCzinc finge
7	SP3	9606. ENSPCSp3 transc
8	MGC10955	9606. ENSPCPutative r
9	IFFO1	9606. ENSPCintermedia
10	TET3	9606. ENSPCtet oncoge
11	GLYR1	9606. ENSPCglyoxylate
12	ZNRF1	9606. ENSPCzinc and r
13	ZNF84	9606. ENSPCzinc finge
14	HOXA2	9606. ENSPChomeobox A
15	PDCD6	9606. ENSPCprogrammed
16	ZSCAN16	9606. ENSPCzinc finge
17	GMG10	9606. ENSPCguanine nu
18	ZNF498	9606. ENSPCzinc finge
19	C17orf104	9606. ENSPCUncharacte
20	MED4	9606. ENSPCmethyl-CpG
21	ZSCAN4	9606. ENSPCzinc finge
22	CBX2	9606. ENSPCchromobox
23	ZNRD1	9606. ENSPCzinc ribbo
24	ZBBX	9606. ENSPCzinc finge
25	SARNP	9606. ENSPCSAP domain
26	AIRE	9606. ENSPCautoimmune
27	C2orf74	9606. ENSPCUncharacte
28	ZNRF3	9606. ENSPCzinc and r
29	ZWILCH	9606. ENSPCZwilch, ki
30	ZP3	9606. ENSPCzona pelli
31	PTEN	9606. ENSPCphosphatas
32	ZNHIT2	9606. ENSPCzinc finge
--	--	--

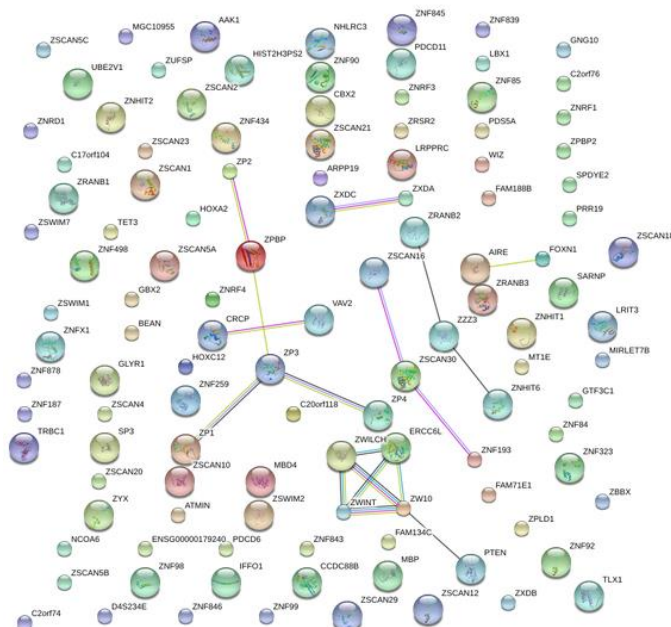
Gene name

Gene annotation

Fig.7. The final filtered list about STRING.

⑪ STRING network analysis report

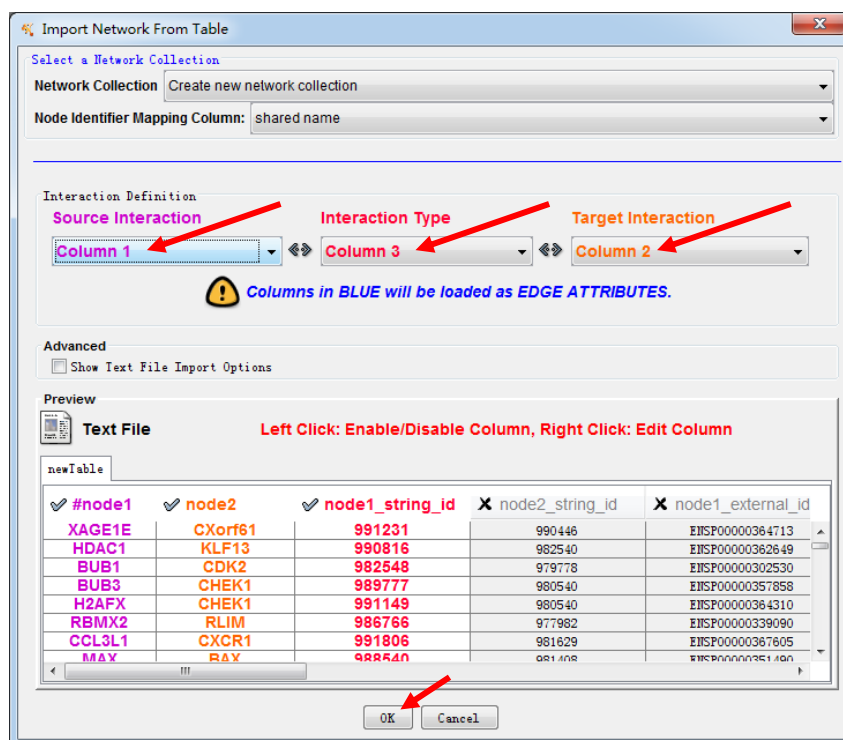
As the result of this analysis, PMD will output PPI-network picture (see fig.8) and the PPI-network text file , which can be read by CytoScape and do further analysis.



biological pathways and integrating these networks with annotations, gene expression profiles and other state data. Additional features are available as Apps. Apps are available for network and molecular profiling analyses, new layouts, additional file format support, scripting, and connection with databases, such as MCODE can find clusters in a network.

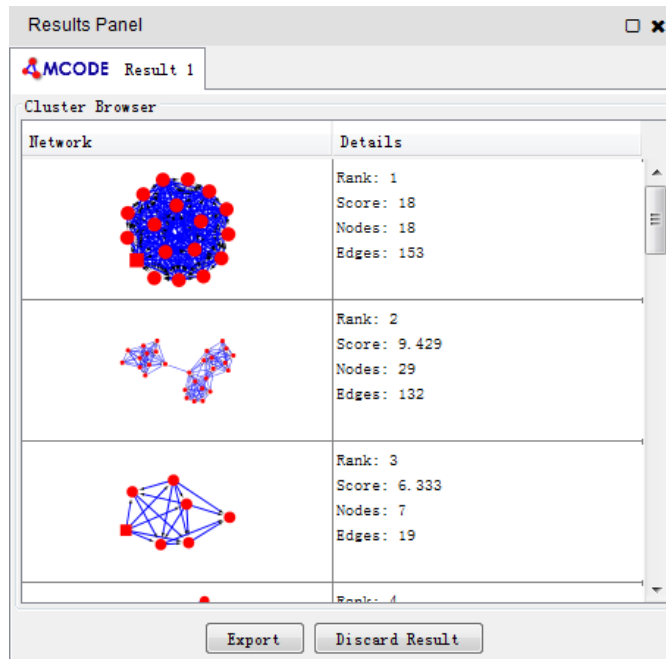
⑬ Cytoscape help document

- A、 You can get available recourses in <http://www.cytoscape.org/download.php>, and install Cytoscape in your computer. (Please install Java 7 first to use Cytoscape)
- B、 Open Cytoscape, and install MCODE.
Apps--App manager--Install Apps--MCODE—Install
- C、 Open PPI-network file in Cytoscape.
File—Import—Network—File...--set up Interaction Definition--OK



D、 Cluster analysis.

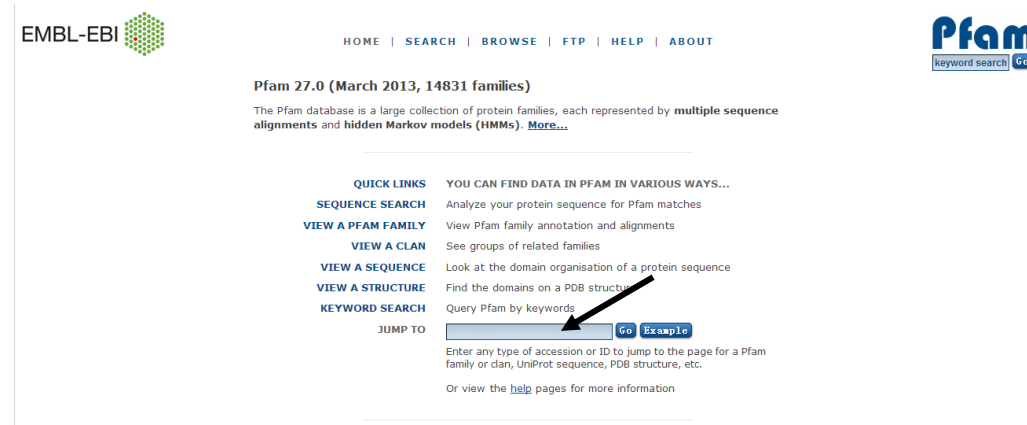
Apps—MCODE—Open MCODE—Analyze current network
And the result will show in the right.



E、 If you want to do further, you can get the complete introduction in http://www.cytoscape.org/documentation_users.html.

⑭ Domain analysis

Proteins are generally composed of one or more functional regions, commonly termed domains. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function. The Pfam database (8) is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models. You can get Pfam ID in hit list, and get more information in <http://pfam.xfam.org/>.



⑮ Motif

Proteins having related functions may not show overall high homology yet may contain sequences of amino acid residues that are highly conserved. Motif analysis will reveal this law.

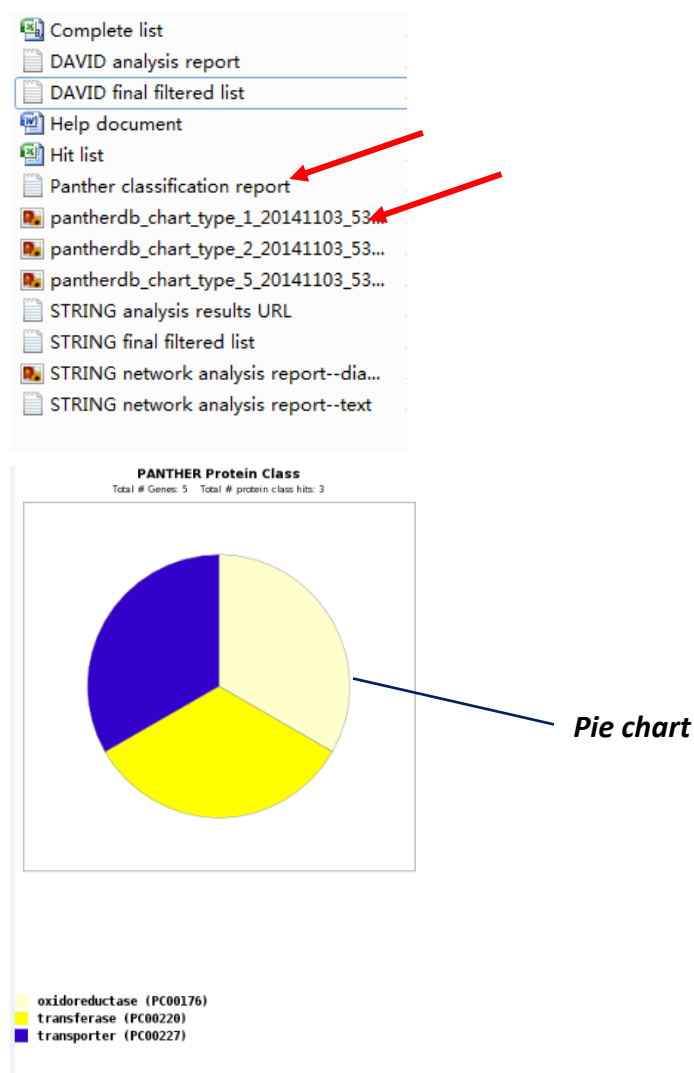
You can get gene sequence in hit list, and submit to MEME(9)

(<http://meme.nbcr.net/meme/cgi-bin/meme.cgi>) or DILIMOT(10)

(<http://dilimot.russelllab.org/>) for doing motif analysis.

⑩ PANTHER

The PANTHER (Protein AnalYsis Through Evolutionary Relationships) Classification System was designed to classify proteins in order to facilitate high-throughput analysis. As the result, pie chart would be downloaded in packet (see fig.). According the result of pie chart, you can search more information in Panther classification report.



⑪ Docking

Docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using, for example, scoring functions.(11)

- Gets the PDB ID in hit list, and download crystal structure in [PDB](#).
- View the crystal structure by molecular visualization software [Pymol](#).
- Using [Autodock Vina](#) for doing molecular docking.

Reference

1. R Development Core Team: R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing 2003.
2. Robert C Gentleman. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* 2004, 5:R80.
3. Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.
4. Da Wei Huang, Brad T. Sherman Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nature Protocols*. 2009;4(1):44-57.
5. Mering Cv. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*. 2003;31(1):258-61.
6. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, et al. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res*. 2005;33(Database issue):D284-8.
7. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*. 2003;13(11):2498-504.
8. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014;42(Database issue):D222-30.
9. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 2009;37(Web Server issue):W202-8.
10. Neduva V, Russell RB. DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res*. 2006;34(Web Server issue):W350-5.
11. Thomas Lengauer, Matthias Rareyt. Computational methods for biomolecular docking. *Curr Opin Struct Biol*. 1996;6(3):402-6.