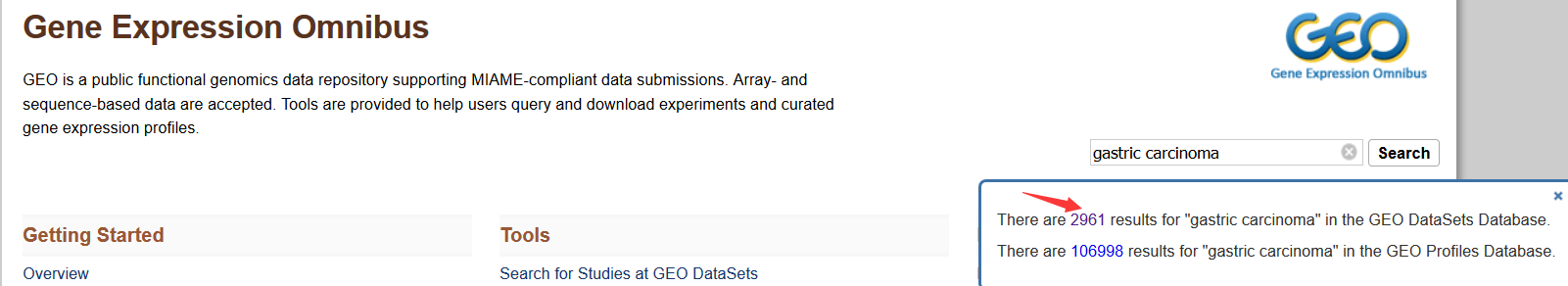
**癌症基因表达数据差异分析**

——精神小伙组

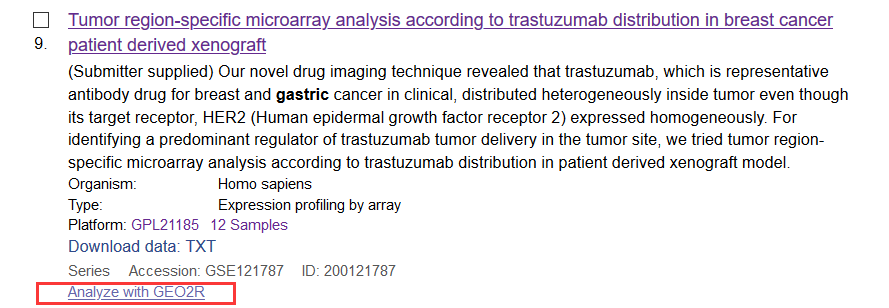
PART1:数据的寻找与下载

进入GEO数据库<https://www.ncbi.nlm.nih.gov/geo/>搜索胃癌基因（gastric carcinoma）

（选其他癌症也可以）



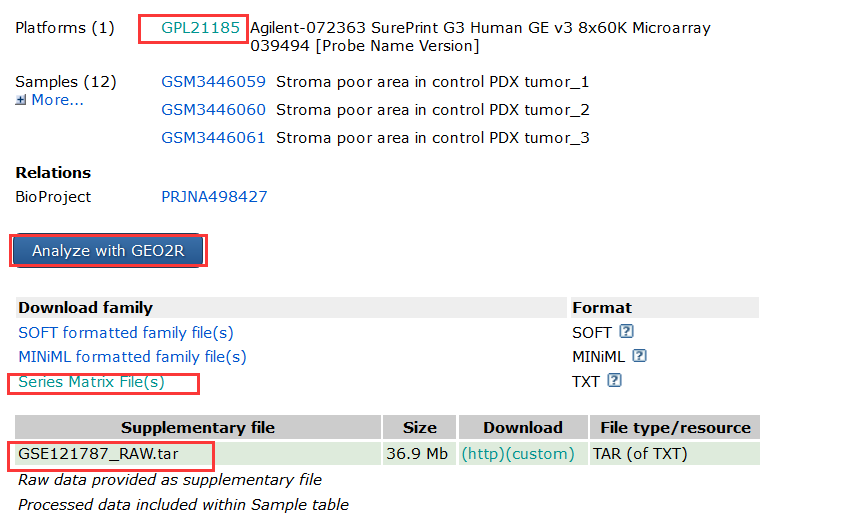
点进去之后，我选择的是，注意：一般选带有Analyse with GEO2R的数据，这些数据提供的GPL探针信息可供下载，GEO数据库也提供线上分析。



点击上图链接

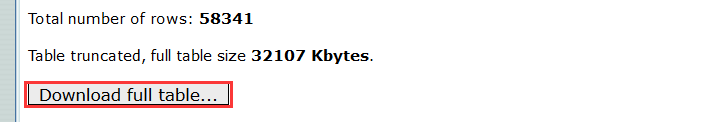


然后得到下图：



我们需要下载的文件有Series Matrix File(s)，GSE121787\_RAW.tar（可选）。

我们还需要的数据是GPL21185的探针文件，点击GPL21185，得到下图：



Download即可。至此，数据下载完毕。

数据解释：

1)GPL是探针文件，里面有着探针ID与基因的相关关系。

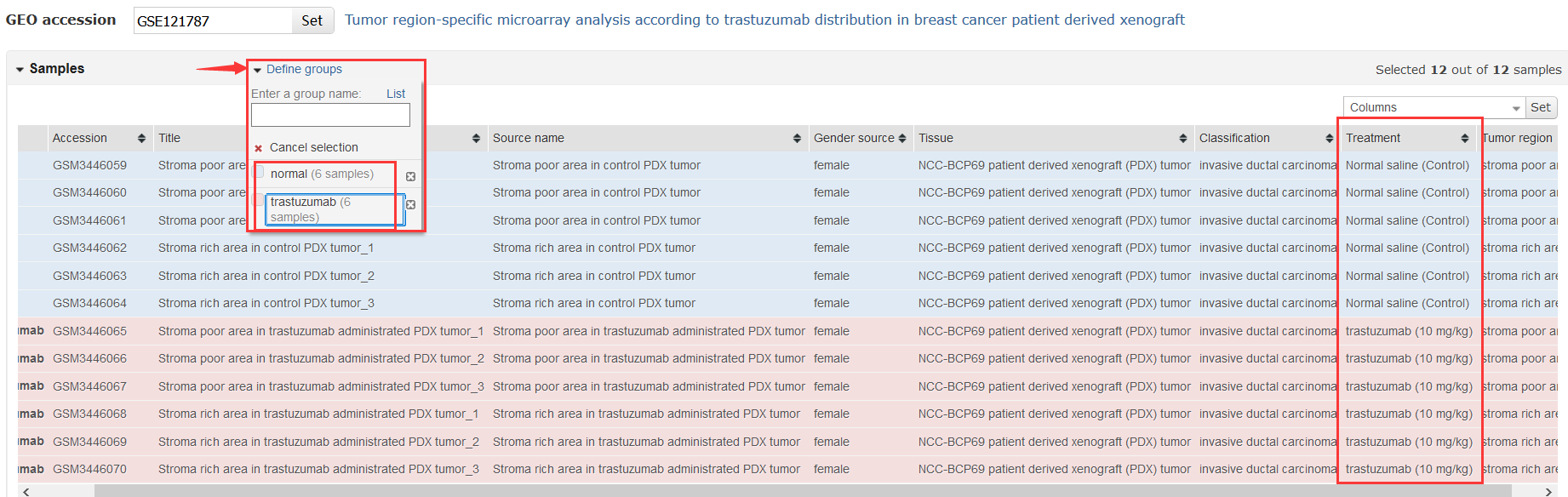
2)Series Matrix File(s)是基因的探针表达矩阵，可以理解为不是真正的基因表达矩阵。因为由基因与探针的关系可知，一个基因可能对应这多个探针。而我们要得到的是行为基因名，列为样本数的基因表达矩阵，因此要对数据进行预处理。

PART2:基因的差异表达分析与数据处理

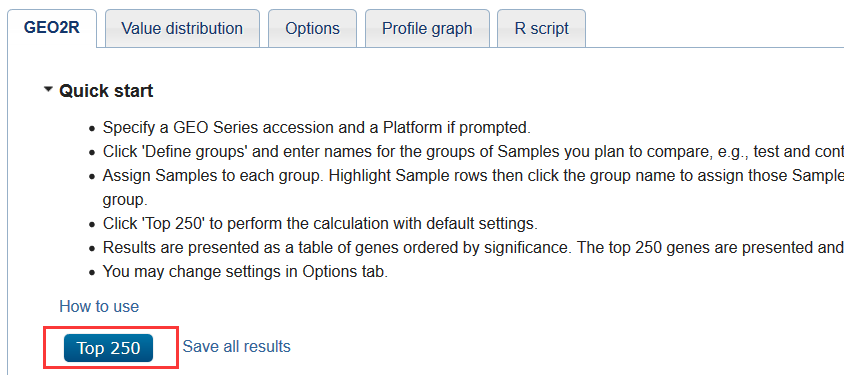
基因的差异表达分析

用GEO数据库GEO2R进行基因的差异表达分析，具体步骤入下：

点击上图的，得到下面的页面，然后给样本进行分组（Define groups），按照正常样本与肿瘤样本进行分组。具体操作如下图所示：



然后点击Top 250，耐心等待即可。

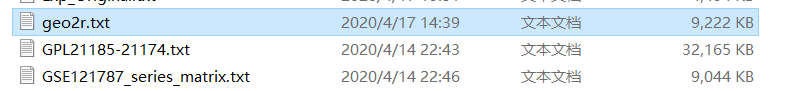


结果如下：



保存为txt文件即可，至于这些字段名代表什么意思建议利用搜索引擎。

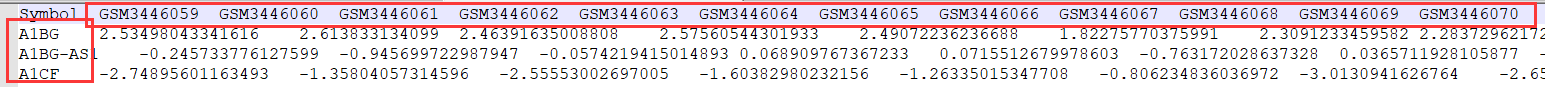
至此，我们已有三个文件：



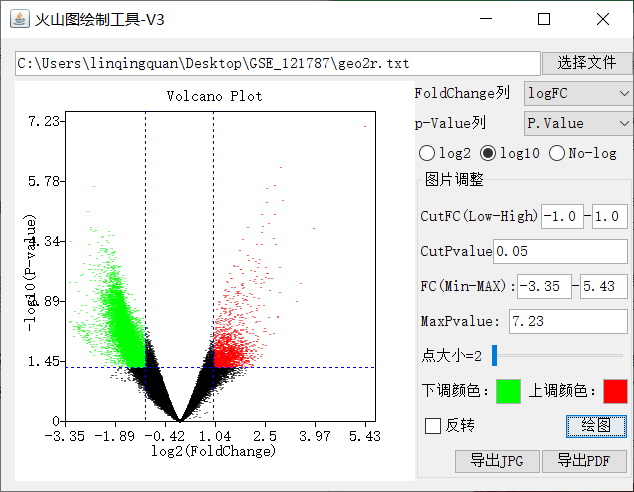
geo2r.txt是GEO2R分析出的结果文件，GPL是探针文件，GSExxx是基因的探针表达矩阵文件。但由于我们最终是要得到差异基因的基因表达矩阵，即行为差异基因名，列为样本数的矩阵，所以我们要对数据进行预处理。

数据预处理

1. 首先用GPL21185-21174.txt和GSE121787\_series\_matrix.txt文件利用R构建一个所有基因的基因表达矩阵（即行为基因名，列为样本数），将其命名为。具体字段为：

构建方法，请在参考文献中参考相应的教程。

1. 由于这个矩阵是所有基因的，我们要在这个矩阵中筛选出差异表达明显的差异基因。首先利用工具SangerBox（须自行下载），绘出geo2r.txt的火山图。SangerBox的下载地址、使用方法，在参考文献中有说明。绘出的火山图如下：

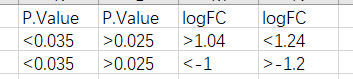


火山图中红色和绿色的区域代表基因差异比较明显。

**火山图的数据解释：**

横坐标是log2（FC），代表倍数变化；纵坐标是-log10(P.value)，P.value代表基因的差异表达程度，越小，基因的差异表达越明显，取-log10之后，相当于值越大表达差异越大

1. 根据火山图给出的信息，我们利用Excel根据-log10(P-value)、log2(FoldChange)的值对geo2r.txt进行筛选。本例中的筛选条件为（列与列之间代表与的关系，行与行之间代表或的关系）：



采用Excel中的高级筛选进行选择

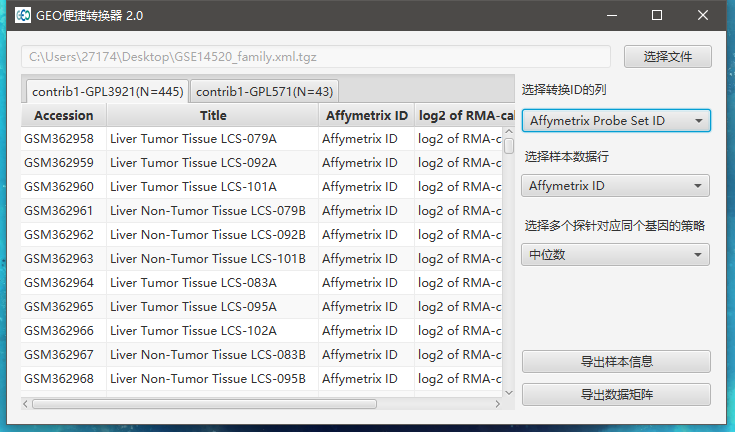


至于怎么进行高级筛选，请参照教程。

将筛选出的差异基因分析文件保存为。

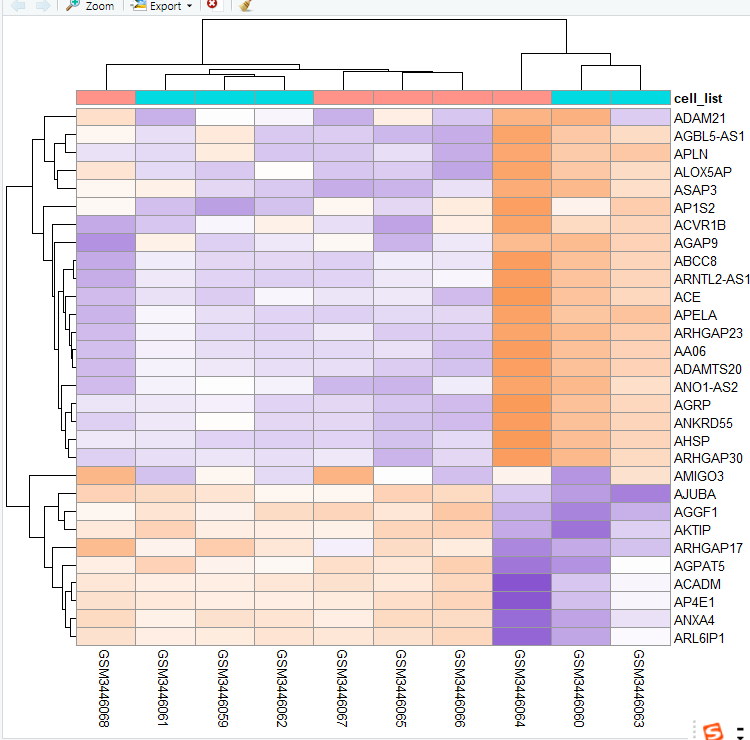
1. 对下载的GPL探针文件利用R取出ID和GENE\_SYMBOL列，与（差异基因分析文件）通过基因探针ID进行合并。最终得到ID-GENE\_SYMBOL的矩阵表。然后根据这个矩阵表与根据ID进行合并。然后删除有空值的行，对行数进行unique处理（防止有相同的GENE\_SYMBOL出现）。最终得到差异基因表达矩阵。以上R语言代码均有参考教程进行详细说明。

PS：在Sangerbox中有一键进行差异表达矩阵生成的工具，只需要从GEO数据库下载基因对应的MINiML文件，导入后可直接生成。参考教程：<https://shengxin.ren/article/135>



PART3：绘制差异基因的火山图和聚类图

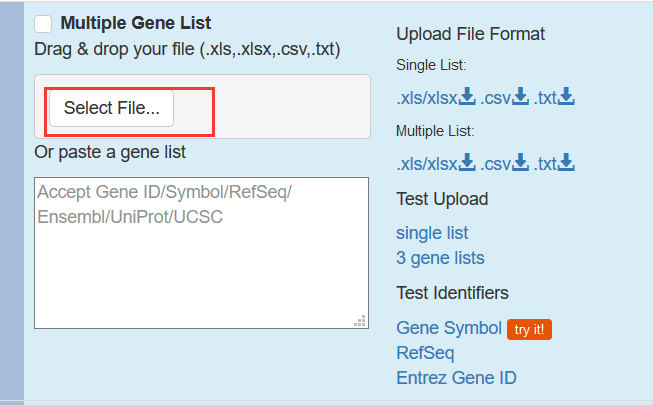
1. 火山图已在上面步骤中汇出，此处不再赘述
2. 用R语言画出聚类热图。结果如下：



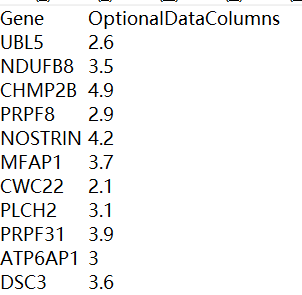
PART4:对差异基因进行GO富集分析和代谢通路分析

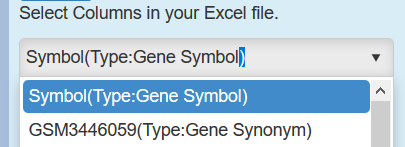
**GO富集分析：**

**方法1：**metascape网站<https://metascape.org/gp/index.html#/main/step1>

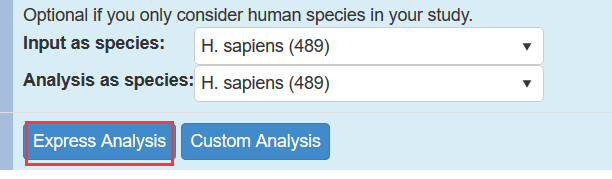


上传基因文件，格式为

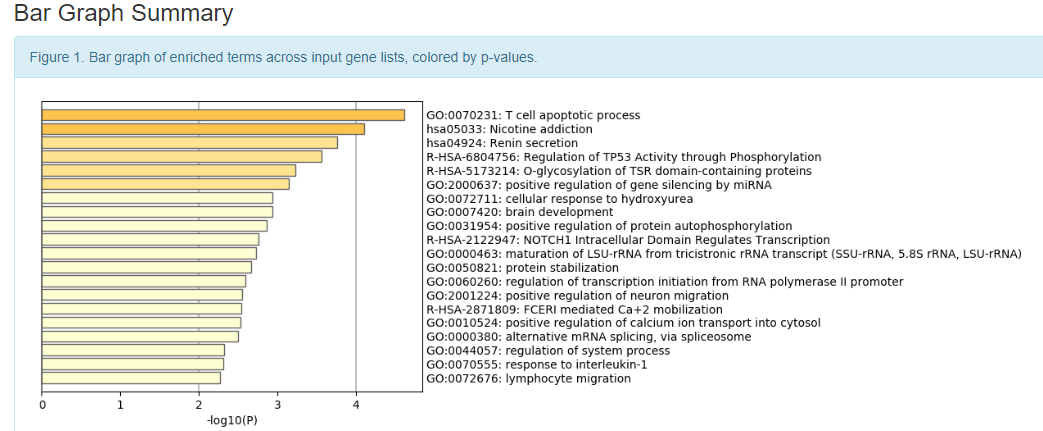


可选列的名字可以任意取，保持一个原则，上传的文件一定带有Gene这一列就可行。对其他字段没有要求。上传本地文件之后，选择基因对应的字段即可。

然后选择物种，此处选择人类

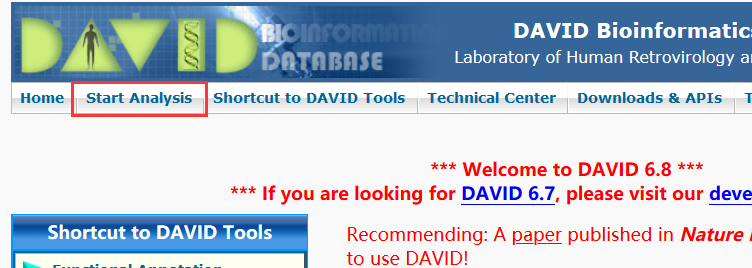


然后点击Express Analysis稍加等待即可。富集分析结果如下：

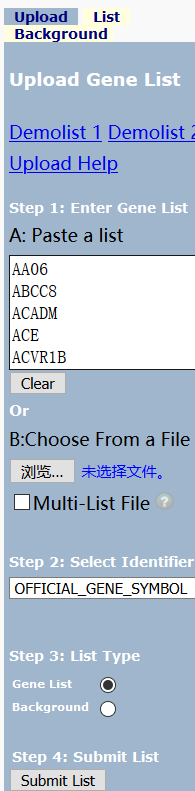


将对应的结果文件下载即可。

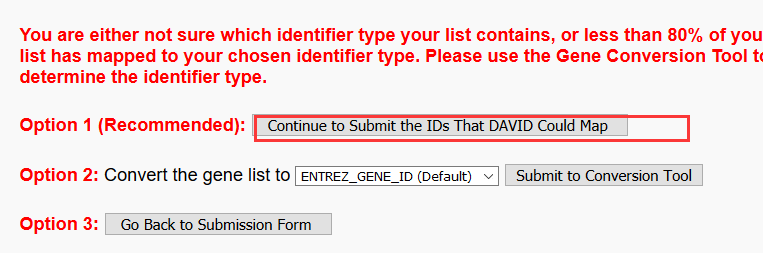
**方法2：**DAVID网站https://david.ncifcrf.gov/



点击红框圈出的地方，Submit List之后耐心等待即可



然后点击continue。

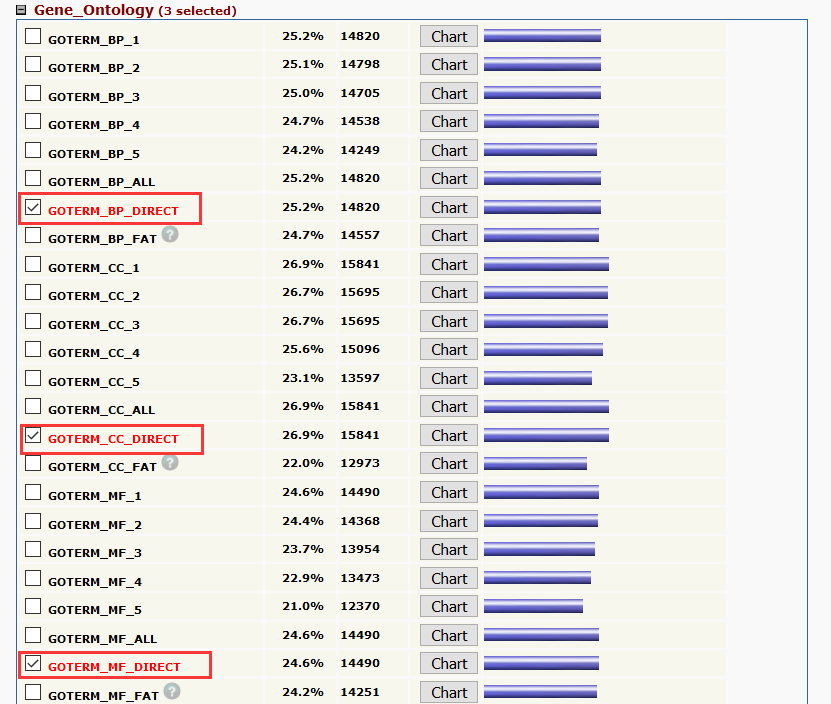


之后

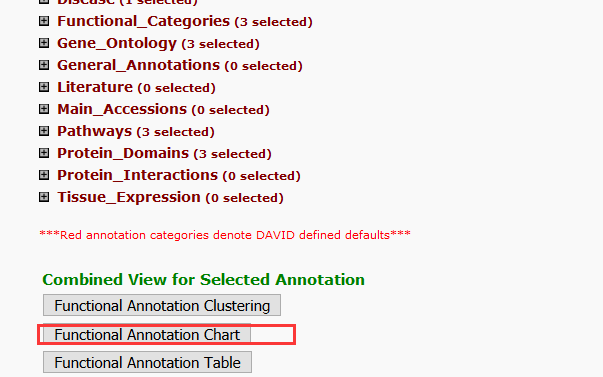


之后首先将默认选项清空。

然后



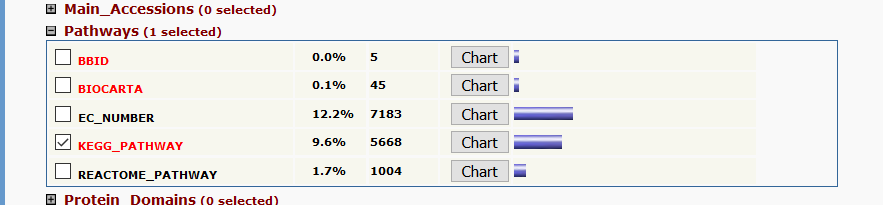
然后



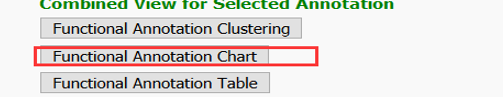
点击即出相应的结果。

代谢通路分析：

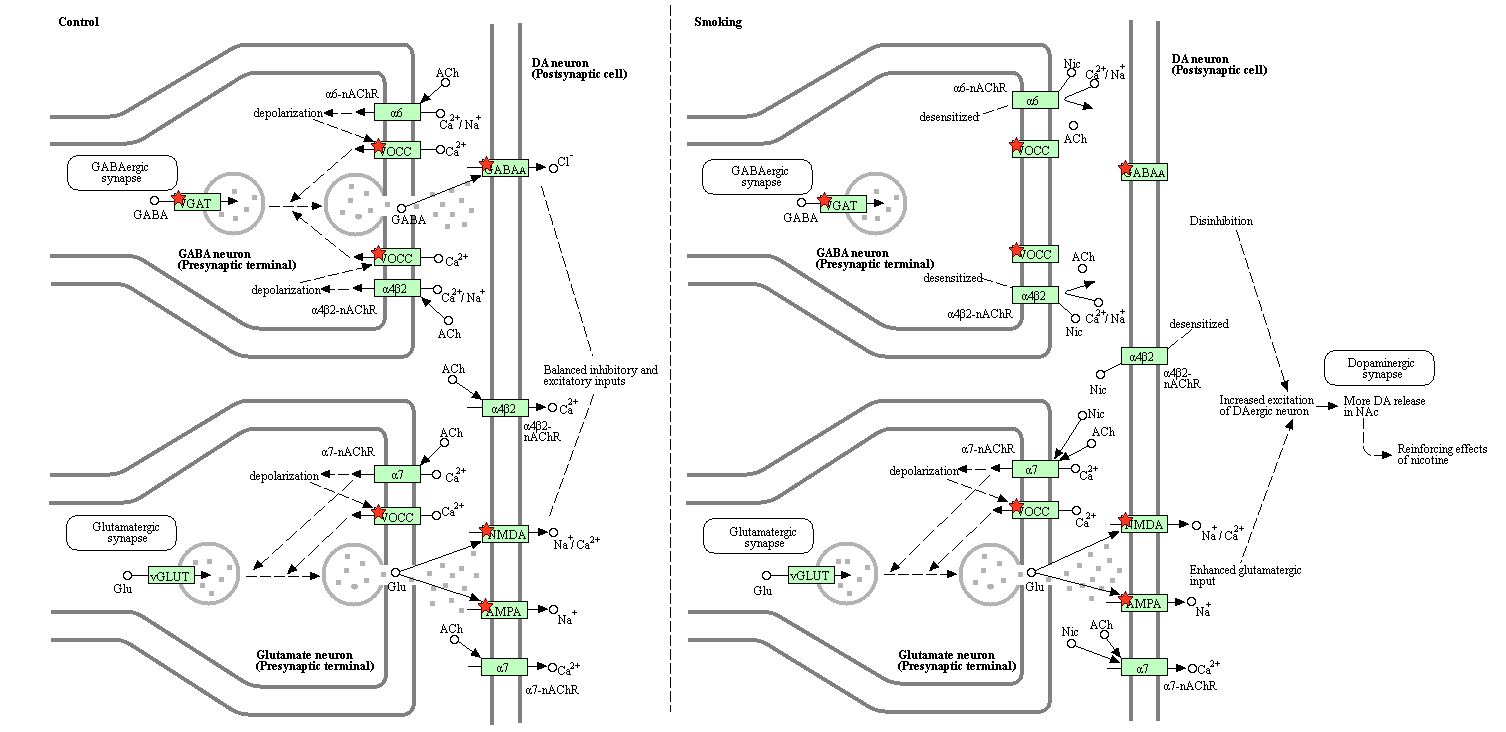
在上述DAVID的基础之上，将默认选项清空之后，点击KEGG\_PATHWAY



之后



即可出每个Term对应的通路图，结果如下：



PART5:Contribution

SZ170110312\_王伟杭：差异基因表达矩阵获取、富集分析、KEGG通路分析

SZ170110314\_林清泉：差异基因表达矩阵获取、火山图绘制、R语言代码、聚类分析、书面报告初稿

SZ170110315\_李江南：ppt制作、presentation

SZ170110331\_刘秋阳：数据的寻找与下载

PART6:参考文献/教程

DAVID的相关使用：

https://jingyan.baidu.com/article/4dc408486c47e0c8d846f159.html

Metascape的相关使用：

<http://www.360doc.com/content/19/0804/09/51784026_852877483.shtml>

火山图绘制以及SangerBox的下载：

<https://www.sohu.com/a/196696514_419916>

真正基因表达矩阵的得到与处理：

<https://www.jianshu.com/p/bc5d885ff4cf>

聚类热图的绘制：

<https://blog.csdn.net/u011262253/article/details/100638123>

本教程中用到的数据合并的R代码：

https://blog.csdn.net/qq\_42754083/article/details/105600999

Excel进行高级筛选：

https://jingyan.baidu.com/article/358570f6819206ce4724fc90.html