1. **背景**

随着全基因组测序和全外显子测序技术的不断发展，为癌症的体突变分析

目前，许多突变分析都是在基因组范围内寻找显著高频突变基因。但是，由于样本的基因突变频率较低，以及基因突变的异质性，传统的分析方法所筛选出的显著突变基因数量较少、假阳性较高。由于基因突变具有异质性，也就是不同个体发生突变的基因不同，但是它们的突变都与癌症的发生发展相关，。基于这一事实，我们以通路为单位，如果通路中有基因发生了突变，我们就判定这条通路发生了改变。这样我们就能计算出每条通路在每个癌症或癌症亚型样本中的发生改变次数。这样，就把以基因为单位的分析转换到以通路为单位的分析，提高了突变的频率。

1. **方法**

2.1、突变文件和通路的获取

我们从firebrowse(http://firebrowse.org)网站中下载各癌型的突变maf数据(注：我们只使用有GENE ID的突变数据，而不管其对应的GENE SYMBOL是否存在，对于突变数据中GENE SYMBOL所对应的GENE ID全为0的癌型，我们使用有GENE SYMBOL的数据)。我们利用KEGG提供的API，从KEGG中获取所有人类的通路名称及通路中的基因，并将与癌症相关的通路去除。每条通路的互作关系是从对应通路图的kgml文件中提取得到（如果是癌(癌亚型)相对于正常的分析，则在提取通路中基因之间的互作关系的同时，去除未突变的基因及与其相关的互作关系）。

2.2、癌(癌亚型)相对于正常的高频突变通路

为了筛选出在癌症样本中显著非随机的高频突变通路，我们提出如下模型：

对于通路i，若样本j(样本量为s)中突变的基因数为，则每次随机从背景参考基因(20635)中取出个基因作为该样本的随机突变基因，判断在随机情况下，通路i在样本j中是否发生突变(只要通路中有一个基因发生突变就判定为该通路发生了改变；有突变：1，不突变：0)。计算出在第r次随机中，通路i在所有样本中的随机突变样本数的，并判断是否大于通路i的实际突变样本数(大于：1，小于：0)。若通路i在n次随机中有m次的随机突变样本数大于实际突变样本数，则通路i的P值为。并对通路P值进行FDR矫正。

2.3、亚型之间的差异突变通路

我们利用Fisher精确检验，筛选出在两类亚型之间突变频率具有显著差异的通路。即，对通路在两类亚型中发生突变的样本数(只要通路中有一个基因发生突变就判定为该通路发生了改变)进行Fisher精确检验，计算如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 腺癌(AD) | 鳞癌(SC) |  |
| 突变 | a | b | a+b |
| 不突变 | c | d | c+d |
|  | a+c | b+d | n=a+b+c+d |

其中a：通路i在亚型1样本中发生突变的样本数，c：不发生突变样本数；b：在亚型2样本中发生突变的样本数，d：不发生突变样本数。利用公式



计算通路i在两类亚型中突变频率的显著性水平，并对通路p值进行FDR矫正。

2.4、子通路提取

2.4.1 连通分量法

对于我们从癌症相对于正常所筛选出的高频突变通路(也可以是感兴趣的通路)，进行子通路的提取。利用从KEGG中提取到的通路互作网络，并将未突变的基因及其所涉及到的互作关系从网络中去除。运用图论中寻找连通分量的方法，从每条通路的互作网络中提取出相互作用的基因模块，作为通路的子通路。

2.4.2 简单路径法

1. **结论**
2. **程序运行介绍**

4.1、第三方包：numpy、pandas、networkx、xlrd、scipy、lxml、cssselect、requests、openpyxl

4.2、进入windows DOS，并将当前路径改为C-Subs程序所在目录

4.3、主要命令为：python C-Subs.py -i [突变maf文件所在目录] -o 结果存放目录 -f 0.05 -n 1000 -s 2 -c 3 -p pathway\_information.xlsx -g gene\_info.xlsx

必须参数：

其中-i:为突变文件所在目录，如果为癌相对于正常的分析，仅需输入该癌型的maf文件所在路径+None；若为两类亚型之间的差异通路分析则需要输入两类亚型maf文件所在路径；

-o:为结果文件所存放的路径；

可选参数：

-f:为FDR控制的截断值；

-n:癌相对与正常的差异通路分析的随机次数；

-s:子通路分析-简单路径法寻找子通路的最小节点数；

-c:子通路分析-连通分量法寻找子通路的最小节点数；

-p:通路信息文件的绝对路径，第一列为通路名称，第二列为通路基因id；

-g:为基因信息文件所在绝对路径，第一列为基因id，第二列为基因symbol。

Example：

1. 癌相对于正常：

命令：python C-Subs.py -i D:/mutation\_data/STAD None -o D:/output -f 0.05 -n 1000 -c 3

1. 亚型之间：
2. 命令：python C-Subs.py -i D:/mutation\_data/LUAD D:/mutation\_data

/LUSC -o D:/output -f 0.05 -s 3