



利用shell编程实现NLP

薛雅文

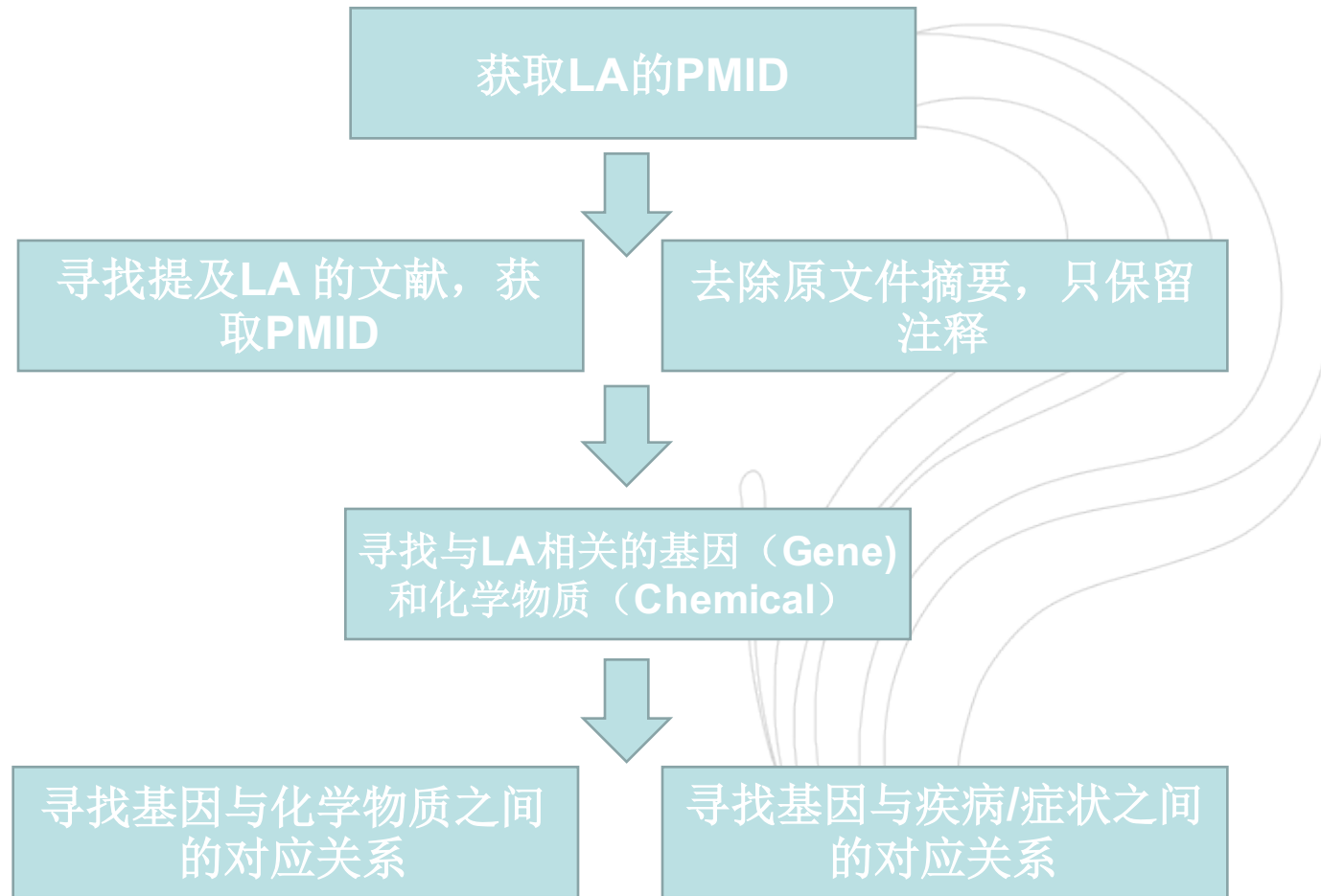
Lung Adenocarcinoma



- **简介：**肺腺癌：是肺癌的一种，属于非小细胞癌。通常起源于支气管粘膜上皮，少数起源于大支气管的粘液腺，发病年龄较小，女性相对多见。
- **发病原因：**
 - 吸烟
 - 大气污染
 - 职业因素长期接触铀镭等放射性物质
 - 肺部慢性疾病如肺结核、尘肺



Workflow





从Pubtator上下载LA的PMID

PubTator

PubMed Lung_Adenocarcinoma Search

Results: 1 to 15 of 12068 Previous Page Next Page Create a collection for this query Add to a new collection

1 [E-cigarettes: voltage- and concentration-dependent loss in human lung adenocarcinoma viability.](#)

Create an annotation collection:

* Collection name:

Lung_Adenocarcinoma

* Collection PMID list:

29665082, 29664061, 29663730, 29662624, 29662612, 29662483, 29662194, 29660690, 29660496, 29660335, 29659171, 29658609, 29657301, 29656868, 29656760, 29656753, 29656749, 29656748, 29656746, 29656111, 29654896, 29651368, 29650826, 29650000, 29649906, 29644529, 29642771, 29642523, 29642152, 29641499, 29637735, 29636879, 29636079, 29634976, 29633541, 29632808, 29632659, 29632545, 29632340, 29631755, 29631611, 29631581, 29631367, 29631033, 29630522, 29629952, 29629530, 29629521, 29628320, 29626120, 29624814, 29623375, 29622850, 29622699, 29622074, 29621876, 29620283, 29620222, 29618796, 29618241, 29617336, 29616327, 29616096, 29615105, 29611055, 29610476, 29608985, 29608981, 29607172, 29606948, 29604585, 29602132, 29600239, 29600083, 29600072, 29600071, 29597147, 29596836, 29596469, 29595658, 29595143, 29593291, 29593202, 29592872, 29590536, 29590374, 29590083, 29588602, 29588590, 29588438, 29588350, 29588349, 29587953, 29587930, 29587911, 29587667, 29581968, 29581811, 29581791, 29581789, 29580760, 29580750, 29580739, 29577897, 29577871, 29577613, 29576870, 29576846, 29576613, 29575939, 29575609, 29575604, 29575527, 29575128, 29574704, 29573824, 29573196, 29572010, 29572008, 29572005, 29572000, 29571999, 29571998, 29571987, 29571784, 29570443, 29569246, 29568376, 29568357, 29567871, 29567476, 29566713, 295665727, 29565453, 29565268, 29565054, 29564776, 29564534, 29563911, 29563587,

e.g., 17317680 16158176 14656948 17187413


寻找提及LA的文献，提取PMID



29660690|t|New 1,2,4-triazole-Chalcone hybrids induce Caspase-3 dependent apoptosis in A549 human lung adenocarcinoma cells. A series of novel 1, 2, 4-triazole/chalcone hybrids was prepared and identified with different structures. The prepared compounds showed remarkable cytotoxic activity against different cancer cell lines. Compounds 24, 25 and 26 showed the highest cytotoxicity among the tested compounds against human lung adenocarcinoma A549 cells with IC₅₀ of 16.04 μM compared to cisplatin with IC₅₀ of 15.3 μM. Flow cytometric analysis of the tested compounds demonstrated that they induced apoptosis in a dose-dependent manner. The further mechanistic study demonstrated that these hybrids induced apoptosis via increased level of proapoptotic protein Bax, release of cytochrome c from mitochondria and activation of caspase-3/8/9 proteins. However, general caspase inhibition by the pan-caspase inhibitor, z-VAD-fmk, significantly inhibited the tested hybrids, suggesting dependency of apoptosis on activation of the caspase-3 pathway.

```
29660690 4 27 1,2,4-triazole-Chalcone Chemical
29660690 43 52 Caspase-3 Gene 836
29660690 81 86 human Species 9606
29660690 87 106 lung adenocarcinoma Disease C538231
29660690 132 148 1, 2, 4-triazole Chemical C045575
29660690 149 157 chalcone Chemical D002599
29660690 313 319 cancer Disease D009369
29660690 4 27 1,2,4-triazole-Chalcone Chemical
29660690 43 52 Caspase-3 Gene 836
29660690 81 86 human Species 9606
29660690 87 106 lung adenocarcinoma Disease C538231
29660690 132 148 1, 2, 4-triazole Chemical C045575
29660690 149 157 chalcone Chemical D002599
29660690 313 319 cancer Disease D009369
```

```
Command: grep -i 'lung adenocarcinoma' laresult.txt | awk '{print $1}'
| sort -n | uniq | awk -F '|' '{print $1}' | uniq > laID.txt
```



```
81307
139205
140093
155940
278851
294062
373915
422569
479125
636378
667821
722727
903615
943561
947267
954007
991153
1159836
1185808
1248011
1280979
1287141
1294457
```



去除原文件摘要，只保留注释

```
29660690 4 27 1,2,4-triazole-Chalcone Chemical
29660690 43 52 Caspase-3 Gene 836
29660690 81 86 human Species 9606
29660690 87 106 lung adenocarcinoma Disease C538231
29660690 132 148 1, 2, 4-triazole Chemical C045575
29660690 149 157 chalcone Chemical D002599
29660690 313 319 cancer Disease D009369
29660690 386 398 cytotoxicity Disease D064420
29660690 434 439 human Species 9606
29660690 440 459 lung adenocarcinoma Disease C538231
29660690 537 546 cisplatin Chemical D002945
29660690 759 784 1, 2, 4-triazole-chalcone Chemical
29660690 855 858 Bax Gene 581
29660690 871 883 cytochrome c Gene 54205
29660690 920 929 caspase-3 Gene 836
29660690 1010 1019 z-VAD-fmk Chemical
29660690 1146 1155 caspase-3 Gene 836
29660496 65 73 tyrosine Chemical D014443
29660496 112 116 EGFR Gene 1956
29660496 125 144 lung adenocarcinoma Disease C538231
29660496 146 157 Lung cancer Disease D008175
29660496 197 204 cancers Disease D009369
29660496 206 214 Patients Species 9606
29660496 256 288 epidermal growth factor receptor Gene 1956
29660496 290 294 EGFR Gene 1956
29660496 307 311 EGFR Gene 1956
29660496 312 320 tyrosine Chemical D014443
29660496 413 421 patients Species 9606
```

Command: `grep '^([0-9]){8}\[^\]' laresult.txt > ref.txt`



寻找与LA有关的基因和化学物质

```
#!/bin/bash
# find gene about LA
echo -e "data pre-processing"

grep -i 'lung adenocarcinoma' laresult.txt | awk '{print $1}' | sort -n | uniq | awk -F '|' '{print $1}' | uniq > laID.txt
grep '^([0-9]{8})[^\|]' laresult.txt > ref.txt

echo -e "data pre-processing done!"

F_id="laID.txt"
F_ref="ref.txt"

echo -e "\n start finding gene and chemical\n"

while IFS= read -r line
do
    grep $line $F_ref | grep 'Gene' | awk -F '\t' '{print $4}' >> gene.txt
    grep $line $F_ref | grep 'Chemical' | awk -F '\t' '{print $4}' >> chemical.txt
done < "$F_id"

echo -e "finish finding!"
echo -e "start remove duplicates"

cat gene.txt | sort | uniq > uniq_gene.txt
cat chemical.txt | sort | uniq > uniq_chemical.txt

echo "finish remove!"
"findGeneandChemical.sh" 28L, 737C written
```

提取Gene和Chemical

排序, 去重



结果文件

uniq_gene.txt

```
14-3-3 isoforms (theta, epsilon, and sigma) and annexin A5
14-3-3 sigma
14-3-3zeta
15-hydroxyprostaglandin dehydrogenase
15-Hydroxyprostaglandin dehydrogenase
15-lipoxygenase
15-PGDH
37-kDa laminin receptor
37-kDa laminin receptor precursor
37LRP
3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase
3'-phosphoadenosine 5'-phosphosulfate synthase 1
-3-phosphoinositide-dependent protein kinase-1
4EBP1
4E-BP1
53BP1
5-lipoxygenase
78 kDa glucose-regulated protein
78-kDa glucose-regulated protein
8-oxo-guanine DNA glycosylase
8-oxoguanine DNA glycosylase
8-Oxoguanine DNA glycosylase
8-oxo-guanine glycosylase-1
```

uniq_chemical.txt

```
10-epi-7,10-epoxy-ar-bisabol-11-ol
10-epi-kalihinol X
10-formyltetrahydrofolate
10-hydroxy-aplysin
10-hydroxycamptothecin
10-hydroxy-debromoepiaplysin
10-hydroxy-epiaplysin
10-methoxy-7-methyl-2H-benzo[g]chromen-2-one
10<sup>-6</sup>
1,10-phenanthroline
1,10-Phenanthroline
11(13)-triene-6,12-olide
1,12-epoxybenz[alanthracene
1,1'-(3,3'-dimethoxybiphenyl-4,4'-diyl)bis(thiourea)
(11)C
(11)C-acetate
11C-acetate
11C-Choline
(11)C-PDT
1,1' dioctadecyl-3-3-3',3'-tetramethylindotricarbocyanine iodide
```


寻找基因与化学物质之间的关系



```
#!/bin/bash
```

```
echo -e "start finding"
while IFS= read -r line1
do
  echo "gene: $line1" >> gc_relationship.txt
  grep "$line1" ref.txt | awk '{print $1}' | sort | uniq > tmp.txt
  while IFS= read -r line2
  do
    grep "$line2" ref.txt | grep 'Chemical' | awk -F '\t' '{print $4}' >> gc_relationship.txt
  done < "tmp.txt"
done < "uniq_gene.txt"
```

```
echo -e "done!"
```

```
~
~
~
~
~
```

```
gene: 14-3-3 isoforms (theta, epsilon, and sigma) and annexin A5
nitrosamine
nitrosamine 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone
NNK
1,4-phenylenebis(methylene)selenocyanate
alpha-1-antitrypsin
NNK
gene: 14-3-3 sigma
manganese superoxide
gene: 14-3-3zeta
BH3
gene: 15-hydroxyprostaglandin dehydrogenase
NAD
phorbol ester
phorbol 12-myristate 13-acetate
PMA
PGE2
sodium butyrate
apicidin
oxamflatin
NAD
prostaglandins
Flurbiprofen
indomethacin
(R)-flurbiprofen
```

□ Gene和Chemical唯一的联系是
PMID

□ tmp.txt: 全部提及该gene
的文献的PMID



```
#!/bin/bash
echo -e "start finding!"
while IFS= read -r line1
do
    echo "chemical $line1" >> dc_relationship.txt
    grep "$line1" ref.txt | awk '{print $1}' | sort | uniq > tmp.txt
    while IFS= read -r line2
    do
        grep "$line2" ref.txt | grep 'Disease' | awk -F '\t' '{print $4}' >> dc_relationship.txt
    done < "tmp.txt"
done < "uniq_chemical.txt"

echo -e "done!"
```

```
chemical 10-epi-7,10-epoxy-ar-bisabol-11-ol
Cancers
lung adenocarcinoma
stomach cancer
hepatoma
colon cancer
cytotoxicity
chemical 10-epi-kalihinol X
Hainan sponge Acanthella sp
lung adenocarcinoma
chemical 10-formyltetrahydrofolate
Cancers
tumors
cancer
cancers
cancer
lung adenocarcinomas
tumor
hepatocellular carcinoma
tumor
chemical 10-hydroxy-aplysin
cancer
lung adenocarcinoma
stomach cancer
hepatoma
```



Thanks !