候选基因信息批量注释软件 v1.0

说明书

2024.1.9

概述：在分析组学数据后，拿到的候选靶基因需要进行进一步的人工筛选，但是大部分基因对于研究者来说比较陌生，需要逐个在NCBI网站通过Gene数据库查询介绍信息，过多的基因游览器查询，将加大研究者花费时间，因此候选基因信息批量注释软件v1.0能够一次性对候选基因进行批量注释，减少研究人员对基因筛查负担。

软件流程：通过Windows系统的DOS命令CURL函数抓取NCBI网站中的靶标基因，使用R脚本对抓取内容进行解析并提取出基因关键信息（包括基因概述、同义词、 染色体位置、基因全名、基因类型等信息），通过在线抓取能够拿到基因最新的描述信息。

使用环境：任意Windows系统，包含DOS命令CURL功能（一般系统自带），R运行环境（推荐4.3.2版本），环境变量配置curl和Rscript，程序自带R-4.3.2安装包及Rscript.exe程序。

使用方法：

1. 准备待查询的基因名文件.CSV格式，列名为Genes并将基因添加到该列下，样例文件为example.csv

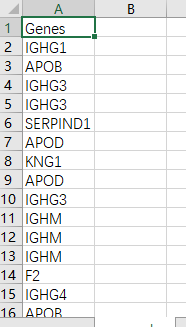


图1 预查询的基因样例文件格式

1. 双击start.bat开始运行程序，选择输入对应功能开始程序运行（输入1为运行程序，输入2为查看教程，输入3为退出程序）





图2 程序主界面

1. 输入1之后运行主程序并输入包含基因名的.CSV文件名按回车键开始运行

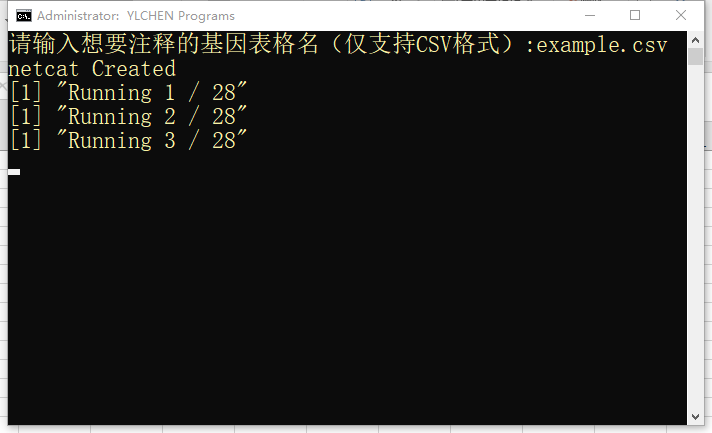


图3 主程序运行界面

1. 运行之后将生成netcat文件网络数据库和名为genesummary.csv的结果文件，通过该.CSV文件可以查看到需要查询的基因功能及描述等信息

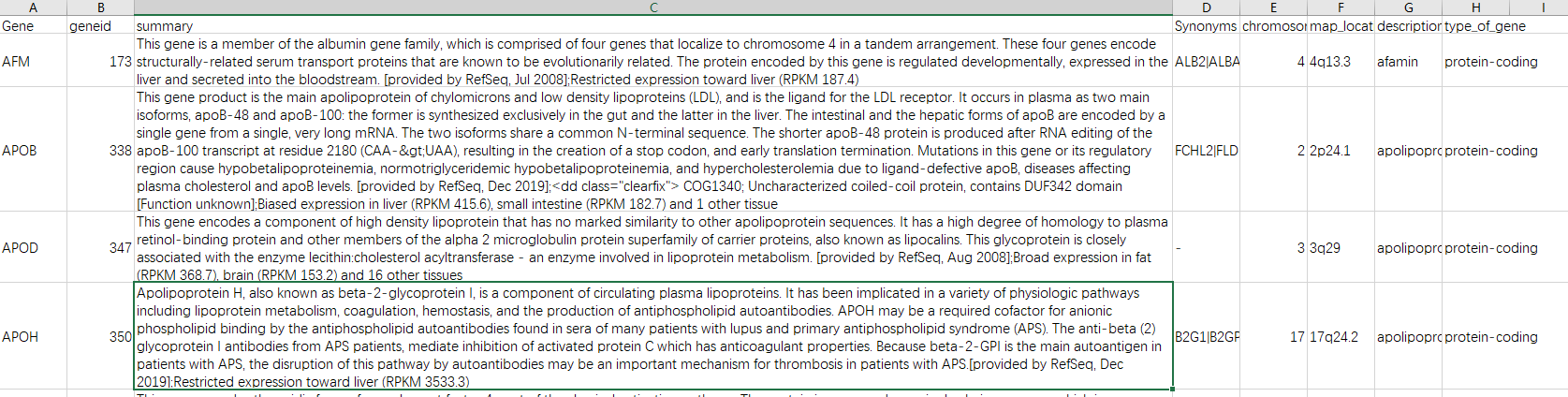


图4 查询结果汇总

其他说明：

1. 基因首次查询会将信息抓取并存入netcat文件夹中，已基因ID命名，如需更新查询删除netcat文件夹即可；
2. 基因ID为具有更新添加的可能，基因ID文件Homo\_sapiens.gene\_info可自行下载更新（1~2年内一般对查询结果影响不大），更新下载地址为https://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE\_INFO/Mammalia/Homo\_sapiens.gene\_info.gz（人种）、其他物种基因查询，比如小鼠需要类似人的Homo\_sapiens.gene\_info文件，并且基因名按照要求填写（人类为全大写字母，小鼠为首字母大写其余小写，其他物种一般与人类一致，特殊物种基因名命名规则需要参考NCBI的命名情况）
3. 其他物种基因查询，需要修改Stargeneanno.bat脚本中的第7行代码中的"Homo\_sapiens.gene\_info"为对应的物种的基因ID信息名称。
4. R环境变量已配置可删除Stargeneanno.bat脚本的第5行和第5行命令。