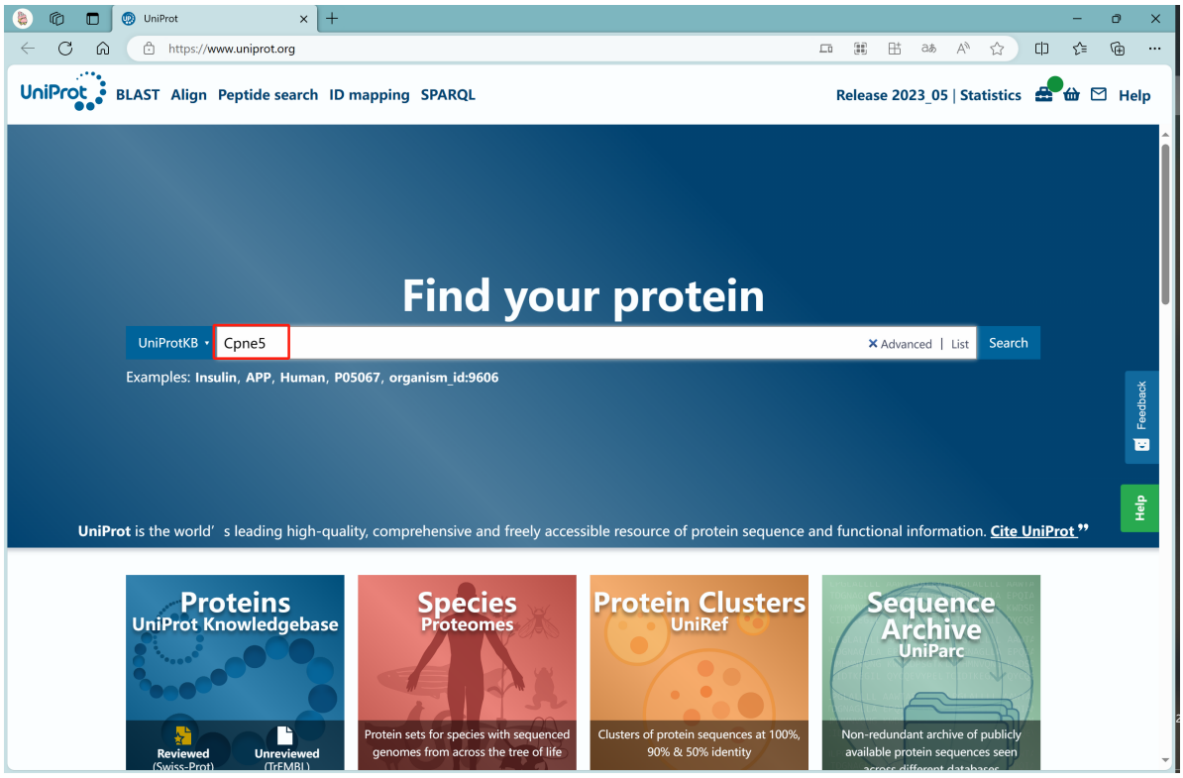
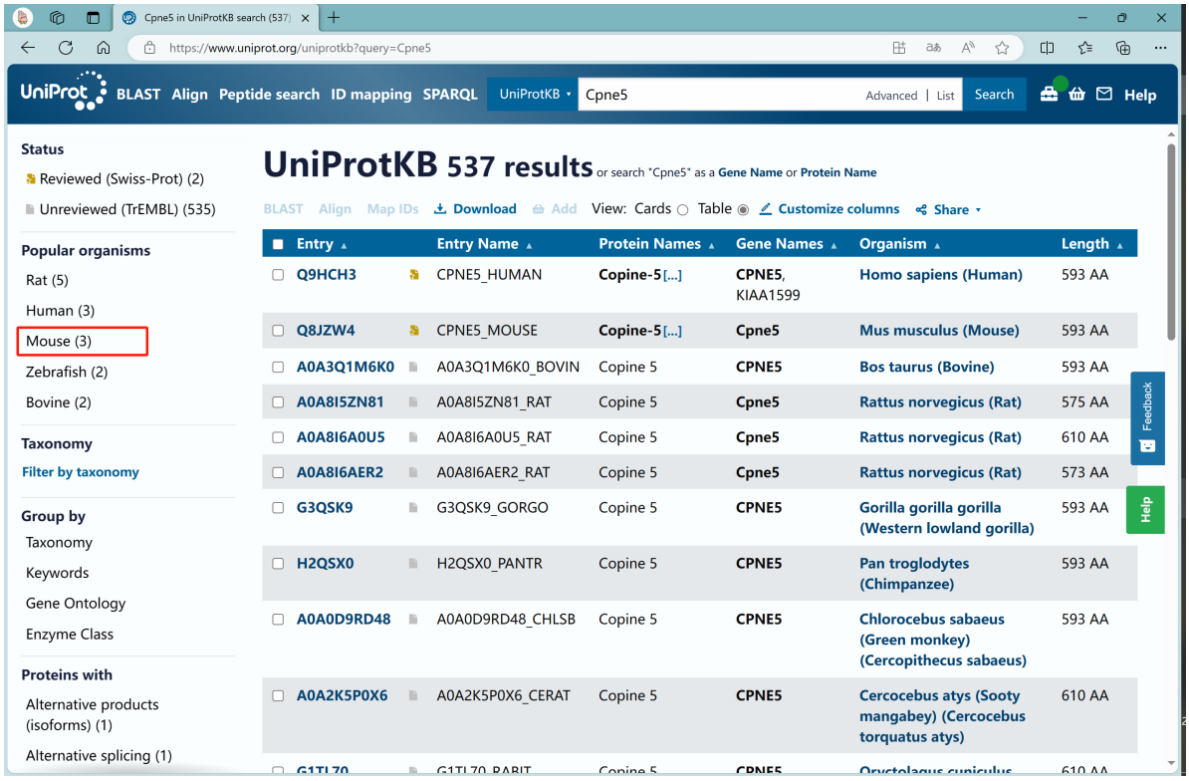
这部分教程将指引你如何获得基因的 ESM-2 编码

查询

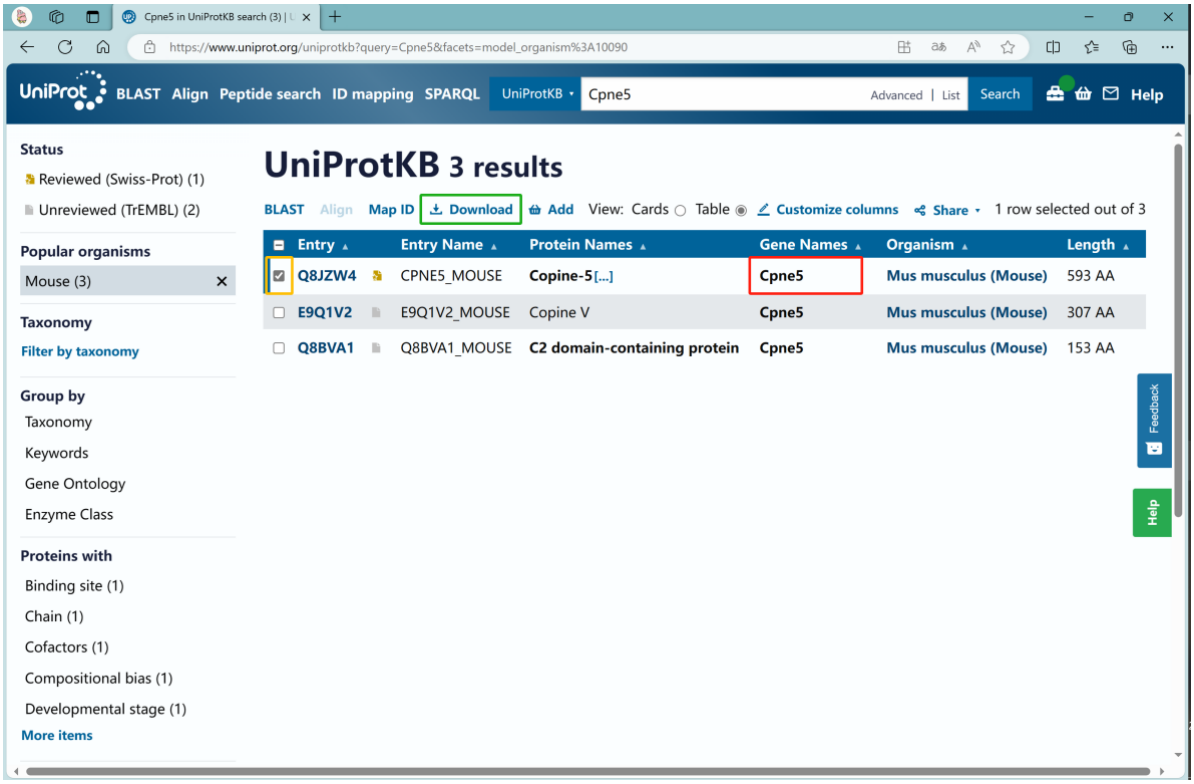
进入 UniProt 网站后，在搜索框搜索基因名（这里以 Cpne5 为例），然后回车：



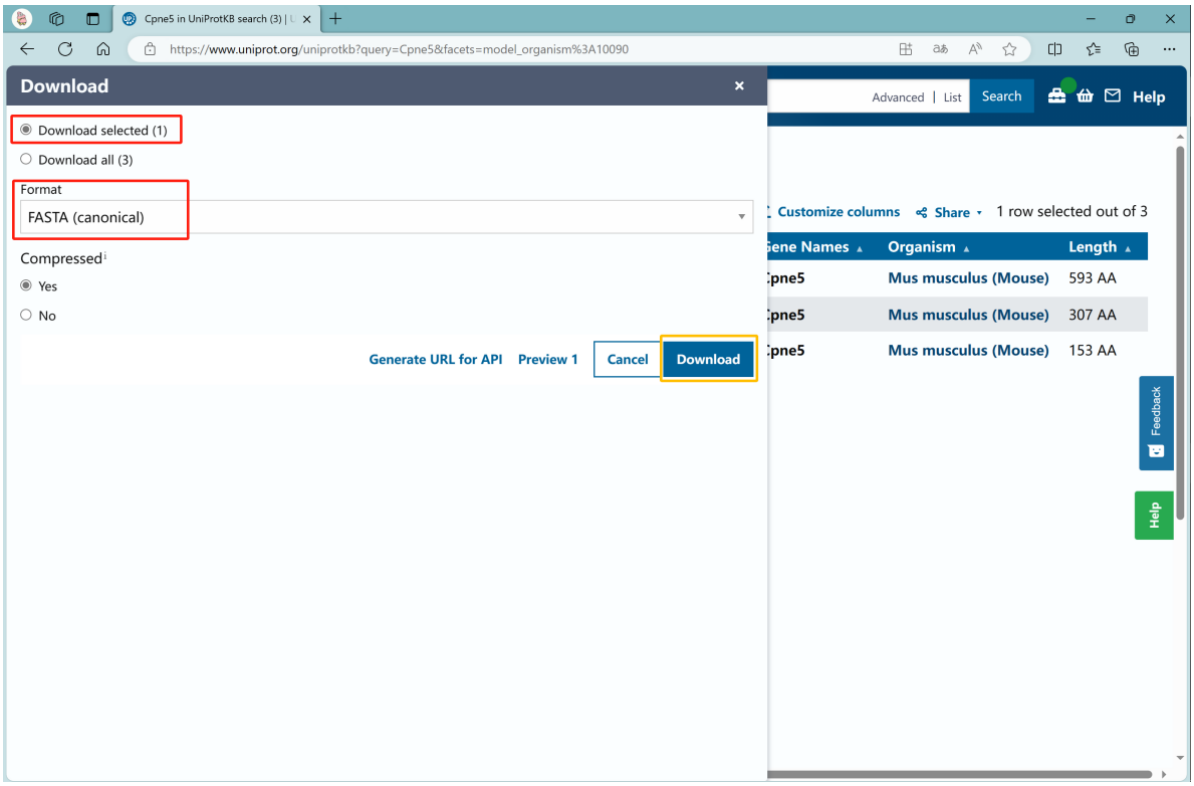
然后在左边筛选框点 Mouse：



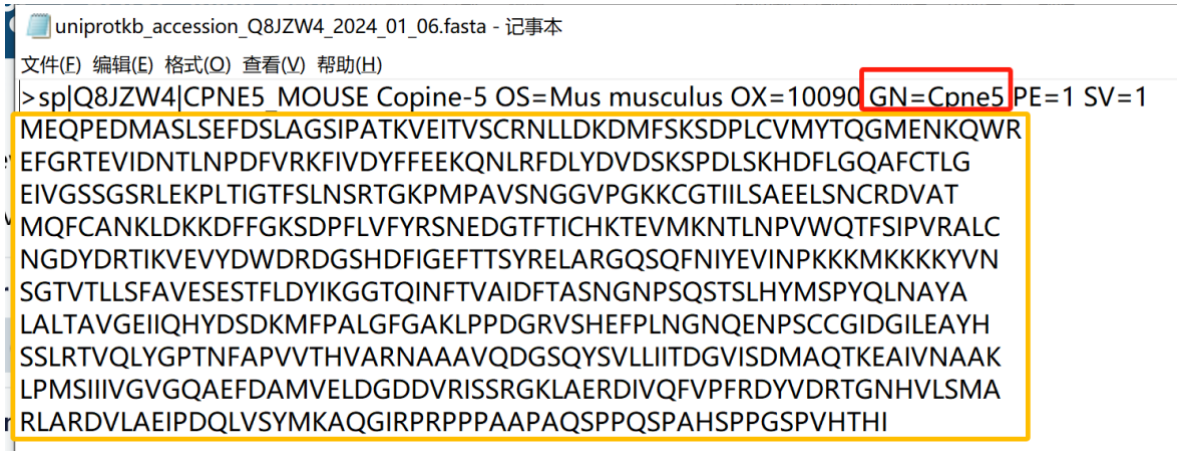
在确定红色框为 Cpne5 的情况下，点击黄色框的选中（Gene Names 相同的情况下选第一个），然后点绿色框的下载；



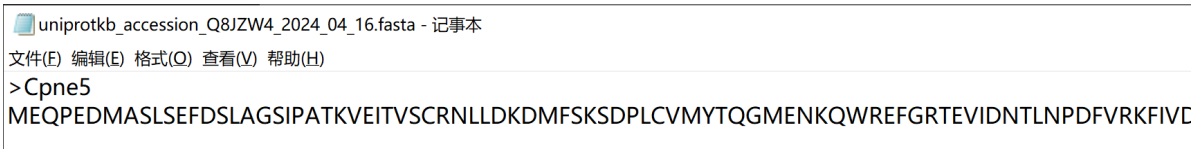
在确定红色框为下载选中的 1 个基因的 FASTA 格式（一般默认就是这样）后，点击黄色框下载



下载并解压后，得到这样的一个文件，其中红色框部分为基因名，黄色框部分为蛋白表达。



将基因名和蛋白表达提取出来后，得到下述文件：



ESM-2 编码

安装好 ESM-2（安装过程可参考 ESM-2）后，输入如下指令：

```
cd esm-main/

python scripts/extract.py esm2_t36_3B_UR50D
uniprotkb_accession_Q8JZW4_2024_04_16.fasta examples/data/some_proteins_emb_esm2
--repr_layers 36 --include mean per_tok

CUDA_VISIBLE_DEVICES=1 python scripts/extract.py esm2_t36_3B_UR50D
/home/songyj/uniprotkb_accession_Q8JZW4_2024_04_16.fasta
examples/data/some_proteins_emb_esm2 --repr_layers 36 --include mean per_tok
```

其中 uniprotkb_accession_Q8JZW4_2024_04_16.fasta 为上文中处理好的基因名和蛋白表达文件；运行成功后会在 esm-main/examples/data/some_proteins_emb_esm2/ 目录下生成基因的编码文件（在这里为 Cpne5.pt）

把你的数据集的基因全部转化为 *.pt 文件后，将其至于一个文件夹下（如 ./pt/），之后运行以下代码：

```
import os
import torch
import pickle
import pandas as pd

df = pd.DataFrame()

for path, dir_lst, file_lst in os.walk(r'./pt'):
    for file_name in file_lst:
        data = torch.load(open(os.path.join(path, file_name), 'rb'))
        print(data['label'])
        print(data['representations'][36][-1])
        df.insert(df.shape[1], data['label'], data['representations'][36]
[-1].numpy())
```

```
pickle.dump(df, open('emb.pkl', 'wb'))
```

即可生成基因的 ESM-2 编码文件