**地点：教二北510和503**

**时间：早8:00-12:00；下午2:00-5:00**

**自带电脑，每个桌子有插座可充电；现场也有台式机，但是需要自己配置。**

**自带水（5楼没有饮水机，教二西2楼有）。**

1. PPT讲解时间：

5分钟讲解，5分钟提问及回答，2点前提交到交作业的服务器。

如果12点前做完可以提前讲，否则就按照题目顺序。

提交PPT和.py文件

命名格式：第一组_小明_小红_小白_小兰_小新.pptx/.py

2. PPT讲解内容：

1）首页标注小组成员，队伍编号，队伍名称

2）运行的最终结果

3）自认为该题的难点所在及对应的解决方案（代码展示），为代码简化所进行的思考

4）每个成员的关键贡献；贡献度排名

3. 评分主要标准：

1）完成度及运行结果的准确性

2）讲解流畅程度，时间把握情况；回答问题是否恰当准确

3）团队合作情况：每个成员都要对代码有所贡献，不能只贡献讲解或做PPT

4）其他

4. 所有文件都是以‘\t’或者空格作为分割。

5. 积极主动使用文心一言等工具

# 第一题

文件：
1.CAD.part.txt为一份冠状动脉疾病全基因组关联分析的部分数据。
2.ref.txt为单核苷酸多态性(single nucleotide polymorphism, SNP)的参考信息文件。

**CAD.part.txt**

chr：SNP所在染色体
bp_hg19：SNP在染色体上的物理位置信息
effect_allele：效应等位基因
noneffect_allele：非效应等位基因
effect_allele_freq：效应等位基因频率
logOR、se、p-value_gc：GWAS结果的beta值，se值，P值
n_samples：样本量
exome：是否位于外显子
info_ukbb：质控质量系数

| chr | bp_hg19 | effect_allele | noneffect_allele | effect_allele_freq | logOR | se_gc | p-value_gc | n_samples | exome | info_ukbb |
|---|---|---|---|---|---|---|---|---|---|---|
| 15 | 102261096 | G | A | 0.00904 | 0.01844 | 0.05371 | 0.731352 | 276916 | no | 0.96 |
| 7 | 154937482 | T | C | 0.63987 | 0.00812 | 0.00892 | 0.362975 | 307471 | no | 1 |
| 19 | 40773859 | AG | A | 0.93785 | 0.02114 | 0.02141 | 0.323612 | 334439 | no | 0.97 |
| 2 | 14201206 | C | T | 0.94532 | 0.02351 | 0.02025 | 0.245555 | 323294 | no | 0.92 |

**ref.txt**

chr：SNP所在染色体
SNP：SNP编号
bp：SNP在染色体上的物理位置信息

| chr | SNP | bp |
|---|---|---|
| 1 | rs142260830 | 234511621 |
| 4 | rs74584348 | 28104062 |
| 2 | rs7368646 | 178062685 |
| 2 | rs575912200 | 161224690 |
| 7 | rs62481722 | 155602064 |

# 第二题

文件:
1. barcodes.tsv为单细胞细胞标签数据。
2. features.tsv为单细胞基因标签数据。
3. matrix.mtx为单细胞矩阵坐标数据。

barcodes.tsv代表细胞标签，一行代表一种细胞的标签

```
AAACATCGAAGGACACCGACTGGA
AAACATCGACAAGCTAAAGGTACA
AAACATCGACAAGCTAAATGTTGC
AAACATCGACAAGCTAGAACAGGC
```

features.tsv两列分别为基因的ENSG id和对应的基因名

```
ENSG00000000419 DPM1
ENSG00000000457 SCYL3
ENSG00000000460 C1orf112
ENSG00000000938 FGR
ENSG00000000971 CFH
```

matrix.mtx为测序结果，第一列对应的是features.tsv的基因标签，第二列对应的是barcodes.tsv的细胞标签，第三列对应检测到的基因表达量。如matrix.mtx中1 2 2表示第一个基因在第二个细胞中表达量为2。
前三行是注释信息， 第三行表示该次测序得到的总信息，即检测到19827个基因，100个细胞，总表达量214164。

```
%%MatrixMarket matrix coordinate integer general
%
19827 100 214164
1 2 2
1 26 1
1 42 1
```

# 第三题

文件:

1.DEP.txt为抑郁症字段数据

2.ukbref.txt为UKB常见基本信息数据

**DEP.txt**

eid代表每个个体的id;

20***为UKB信息代码

UKB中抑郁症相关字段主要有:

20441：Ever had prolonged loss of interest in normal activities

20446：Ever had prolonged feelings of sadness or depression

20546：Substances taken for depression

其中20441和20446的值有两类，1代表yes, 0代表no。

20546值有三类，1代表Unprescribed medication (more than once)，3代表Medication prescribed to you (for at least two weeks), 4代表Drugs or alcohol (more than once)。 20546-0.1， 20546-0.2，20546-0.3代表进行了三次问卷，至少有一次问卷符合要求即可。

| eid | 20420-0.0 | 20421-0.0 | 20425-0.0 | 20441-0.0 | 20446-0.0 | 20546-0.1 | 20546-0.2 | 20546-0.3 |
|---|---|---|---|---|---|---|---|---|
| 1342736 | | | | | | | | |
| 1684809 | | -121.0 | 0.0 | 1.0 | 1.0 | 3.0 | | |
| 3901488 | | 0.0 | 0.0 | 0.0 | 0.0 | | | |
| 2369214 | | | | | | | | |
| 2518745 | | | | | | | | |

**ukbref.txt**

第一列为eid，第二列为性别，第三列为出生年份，第四列为收录信息日期，第五列为人种信息，人种代码信息具体可参考UKB网站https://biobank.ndph.ox.ac.uk/showcase/coding.cgi?id=1001。

| | | | | |
|---|---|---|---|---|
| 2961594 | 0 | 1943 | 2009-12-15 | 1001 |
| 4374903 | 1 | 1962 | 2009-09-14 | 1001 |
| 4925783 | 0 | 1949 | 2009-09-30 | 1003 |
| 2725586 | 1 | 1951 | 2009-03-28 | 1002 |

# 第四题

文件：
1.METSIM.txt为皮下脂肪组织的部分eqtl数据
2.Gencode.V19.genes.anno为基因参考文件

METSIM.txt
第一、二列为基因ENSG号和基因名
第三列为SNP物理位置+等位基因，即chr:BP_A2_A1
第四列为SNP编号
第五列为等位基因频率
第六、七、八列为P值，beta值，se值

Gencode.V19.genes.anno
第一列为基因所在染色体
第二列为基因转录起始位点（TSS)
第三列为基因转录终止位点（TSE)
第四列为genomic strand
第五列为基因ENSG号
第六列为基因类型
第七列为基因名

```
ENSG00000198862.9   LTN1    21:29365459_C/T rs112674045 0.0576037   0.63889 0.0663119   0.141214
ENSG00000198862.9   LTN1    21:29365909_A/G rs117835687 0.031106    0.553247    -0.120445   0.202987
ENSG00000198862.9   LTN1    21:29366034_G/A rs73189325  0.0218894   0.442461    -0.192614   0.250555
ENSG00000198862.9   LTN1    21:29367527_T/C rs115804016 0.0506912   0.83243 0.0318696   0.150531
ENSG00000198862.9   LTN1    21:29367528_G/A rs114393446 0.0506912   0.833422    0.0316711   0.150498
```

```
chr1    11869   14412   +   ENSG00000223972.4   pseudogene  DDX11L1
chr1    14363   29806   -   ENSG00000227232.4   pseudogene  WASH7P
chr1    29554   31109   +   ENSG00000243485.2   lincRNA MIR1302-11
chr1    34554   36081   -   ENSG00000237613.2   lincRNA FAM138A
chr1    52473   54936   +   ENSG00000268020.2   pseudogene  OR4G4P
chr1    62948   63887   +   ENSG00000240361.1   pseudogene  OR4G11P
chr1    69091   70008   +   ENSG00000186092.4   protein_coding  OR4F5
chr1    89295   133566  -   ENSG00000238009.2   lincRNA RP11-34P13.7
```