宏基因组在复杂疾病中的应用习题说明

某些复杂疾病因为通过传统方法诊断困难，而又因为肠道菌群或者其他部位的菌群与该种复杂疾病相关性较强，因此研究人员会使用与疾病比较相关的某些菌的数据来对疾病进行分类。此次习题需要大家有一定的编程语言（R或者python等）以及统计学习基础。本次习题不限方法，可以通过随机森林、lasso或者其他方法对菌群数据进行筛选并且使用筛选后的菌群数据达到相对较好的疾病分类效果。

表中的数据包括四个表单。因为保证判别模型的稳健性除模型训练样本外，还需要外部验证数据（此处的外部数据不参与训练模型的构建，只用于判断训练模型的效果）。因此本次练习中包括两部分的数据，baseline和validation。各自分别表示基线样本和验证样本。

1. 表单“baseline-group”为基线样本名称以及分组信息。其中HC表示健康人，SCZ表示病人。
2. 表单“baseline-profile”为样本-菌群相对丰度数据，行名为肠道菌的名称，列名为样本名称。
3. 表单“validation-group”为基线样本名称以及分组信息。其中HC表示健康人，SCZ表示病人。
4. 表单“validation-profile”为样本-菌群相对丰度数据，行名为肠道菌的名称，列名为样本名称。

提示：

1. 常用的模型：随机森林，lasso等
2. 模型效果判断：AUC（Area Under Curve）：ROC曲线下与坐标轴围成的面积。该值越接近1越好。
3. 注意过拟合的问题。
4. 注意需要对菌群数据进行筛选，即实现少量的菌也能达到较好的判别效果。
5. R或者python中都有相应的包或者模块，请自行查阅。