# TCGA-LUAD数据集简介

**LUAD数据**存储在超算服务器上，具体存储路径为：/lustre/home/acct-lctest/stu411/LUNG/LUAD。

训练集和测试集按照4：1的比例，其中对应的样本编号在sample_index(train&test).xlsx文件中。

**LUAD文件夹**中包含了TCGA(The Cancer Genome Atlas)数据库中有关肺腺癌(Lung Adenocarcinoma)的临床信息、病理图像以及组学数据，分为如下四个子文件夹。

```
Clinical  Image  ProcessedData  RawData
```

**Clinical文件夹**中包含患者临床相关的一些信息，含有两个txt文件：

```
nationwidechildrens.org_clinical_drug_luad.txt
nationwidechildrens.org_clinical_patient_luad.txt
```

其中nationwidechildrens.org_clinical_patient_luad.txt为临床信息，nationwidechildrens.org_clinical_drug_luad.txt为药物治疗信息。

例如，临床信息文件中共有123列属性，其对应的列名在第一行有相应的解释。包含的主要变量有：

| 变量名 | 定义 | 举例 |
|---|---|---|
| bcr_patient_barcode | TCGA中患者统一编码 | TCGA-05-4245 |
| race | 种族 | white |
| gender | 性别 | female, male |
| age_at_initial_pathologic_diagnosis | 年龄 | |

| | | |
|---|---|---|
| vital_status | 生存状态 | death, alive |
| last_contact_days_to | days_to_last_followup(最后随访时间) | |
| death_days_to | days_to_death(死亡时间) | |
| tobacco_smoking_history_indicator | tobacco_smoking_history | 1, 2, 3, 4, 5 |
| tobacco_smoking_pack_years_smoked | number_pack_years_smoked | continuous variable |
| carbon_monoxide_diffusion_dlco | dlco_predictive_percent | 肺功能指标(continuous) |
| ajcc_pathologic_tumor_stage | pathologic_stage(病理分期) | StageI, StageII, …… |
| icd_10 | icd_10(疾病分类) | |

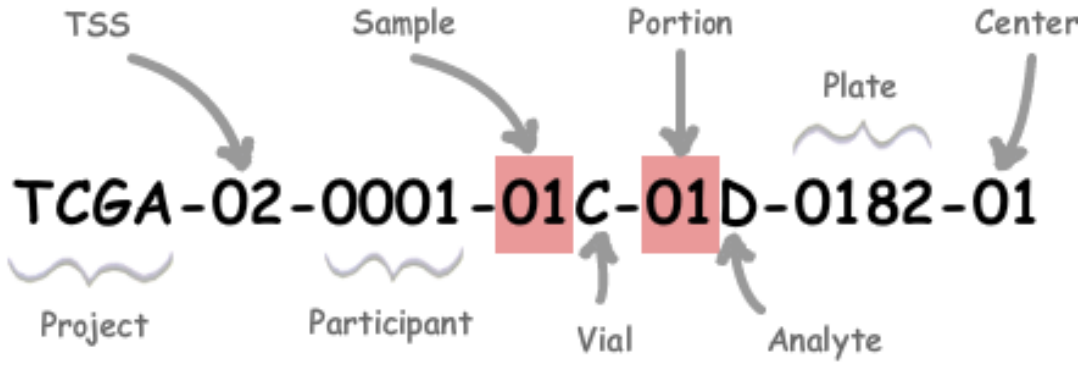
其中，关于tobacco_smoking_history_indicator表示患者的吸烟历史（分类变量），即1(Lifelong Non-smoker (less than 100 cigarettes smoked in Lifetime)); 2(Current smoker (includes daily smokers and non-daily smokers or occasional smokers)); 3(Current reformed smoker for > 15 years); 4(Current reformed smoker for ≤15 years); 5(Current reformed smoker, duration not specified)；tobacco_smoking_pack_years_smoked表示吸烟量(连续型变量)，即Pack-years were defined as the number of cigarettes smoked per day divided by 20 and then multiplied by the number of years smoked。具体详情也可参见clinical_colnames.xlsx。

**Image文件夹**中包含患者肺部的组织病理图像，由H&E方法进行染色，以svs格式储存，这是常见的储存病理图像方式。Image文件又包含1067个子文件（病理图形）。

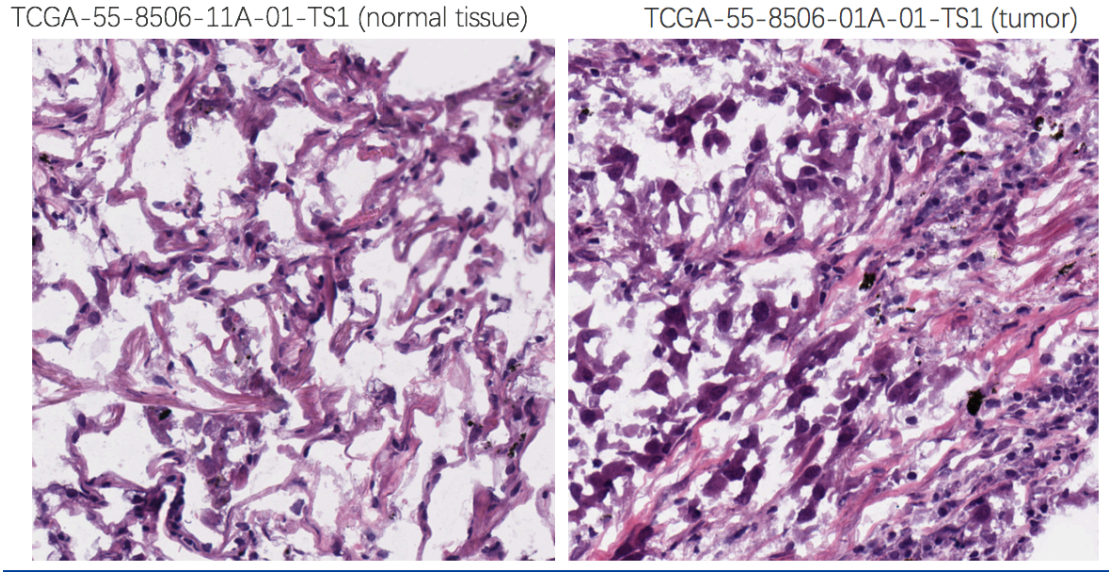
同一个患者可能会有多张病理图像，每一张病理图像都有特定的编码，遵循TCGA的编码格式。在下图svs文件中，TCGA-91-6831为患者编码，

与临床信息中的bcr_patient_barcode相对应。其后的01代表Sample，这两个数字十分关键，编号01~09表示肿瘤，10~19表示正常对照，该位置最常见的就是01和11。A是Vial，表示所选样本在一系列患者组织中的顺序，绝大多数样本该位置编码都是A，少数的是B。01A-之后的01是Portion，表示一个患者组织中不同部分的顺序编号，同一组织会分割为100-120mg的部分，分别进行使用。TS1为组织病理切片编号，TS表示Top Slide，该位置还有BS(Bottom Slide)和MS(Middle Slide)，紧随其后的1代表病理切片的顺序。

```
TCGA-91-6831-01A-01-TS1.4b44cb99-5b5e-4999-bc73-b836b1536d9d.svs
```




举例如下：




**RawData文件夹**中包含TCGA的原始组学数据。**ProcessedData文件夹**中将原数据进行预处理，得到了Matrix Data，存在SNP和RnaFPKM两个子文件夹中。SNP文件夹中的LUAD.Mutation.Sample565.txt为突变数据(LUAD Mutaion Matrix)，每一行为基因，每一列为样本，RnaFPKM 文件夹中RNA.FPKM.Tumor513.txt为表达数据(LUAD Tumor Expression Matrix)，每一行为基因，每一列为样本。