**TMB标准化执行&结果文档**

——依据《国家药品标准物质使用说明书》，《附件2.代码使用说明 360042》

|  |  |
| --- | --- |
| 撰写人 | 朱凤娇 |
| 邮箱 | zhufengjiao@genomics.cn |
| 时间 | 20220211 |

1. 关于TCGA MC3数据获取、TMB计算与线性回归拟合

执行路径：/jdfstj1/B2C\_COM\_P1/5\_Personal/zhufengjiao/TMBstand/SortOut

Python：/zfstj1/BC\_COM\_P01/clinical/5\_Personal/zhufengjiao/miniconda3/bin/python

* 1. 需下载和整理的文件：
     1. **TCGA MC3数据：mc3.v0.2.8.PUBLIC.maf.gz**

(获取方式：<https://gdc.cancer.gov/about-data/publications/mc3-2017>)

* + 1. **TCGA样本项目数据： combined\_study\_clinical\_data.tsv**
    2. **Homo\_sapiens.GRCh37：Homo\_sapiens.GRCh37.87.gtf**

(获取方式：wget <ftp://ftp.ensembl.org/pub/grch37/release->93/gtf/homo\_sapiens/Homo\_sapiens.GRCh37.87.gtf.gz

gunzip -c Homo\_sapiens.GRCh37.87.gtf.gz）

* + 1. **idt芯⽚bed数据：xGen\_Exome\_Research\_Panel\_target.bed**

（获取方式：<http://www.nanodigmbio.com/?app=site&m=download&a=downloadFile&id=76795>）

* + 1. **BGI Panel flank50Bed数据：flk50\_bgi.bed**

（获取方式：cat pan\_cancer.chip.v1.flank50.sort.merge.bed | sed 's/chr//g' > flk50\_bgi.bed）

* + 1. BGI使用驱动基因List：driver.txt
  1. 依据《国家药品标准物质使用说明书》对TCGA数据进行过滤：

Run：/zfstj1/BC\_COM\_P01/clinical/5\_Personal/zhufengjiao/miniconda3/bin/python 1-filterMc3.py

Input: mc3.v0.2.8.PUBLIC.maf; combined\_study\_clinical\_data.tsv

Output：mc3.lite.tsv

* 1. 处理idt，panel的bed：

# ⽣成idt芯⽚上下游FLK50bp后到bed⽂件

Run：sh step2.1.sh

Input: xGen\_Exome\_Research\_Panel\_target.bed

Output: flk50\_idt.bed

# 生成FLK2bp的CDS区间文件：

Run：sh step2.2.sh

Input: Homo\_sapiens.GRCh37.87.gtf

Output: Homo\_sapiens.GRCh37.87.v93\_CDS.flk2.bed

# WES的FLK50区间 与 CDS的FLK2 的交集区间⼤⼩将⽤作TMB计算的分⺟，计算结果为34M

Run：sh step2.3.sh

Input: Homo\_sapiens.GRCh37.87.v93\_CDS.flk2.bed

Output: intersect\_idt\_cds.bed

#依照2-restrainRegions.sh 限制突变区域。该脚本接受⼀个positional argument，即WES的bed⽂件

Run：sh step2.4.sh flk50\_idt.bed

Input: flk50\_idt.bed

Output: mc3.lite.idt\_intersected.tsv

# Panel的FLK50区间 与 CDS的FLK2 的交集区间⼤⼩将⽤作TMB计算的分⺟，计算结果为34M

Run：sh bgistep2.3.sh

Input: flk50\_bgi.bed; Homo\_sapiens.GRCh37.87.v93\_CDS.flk2.bed

Output: intersect\_bgi\_cds.bed

#依照2-restrainRegions.sh 限制突变区域。该脚本接受⼀个positional argument，即panel的bed⽂件

Run：sh bgistep2.4.sh

Input: intersect\_bgi\_cds.bed

Output: mc3.lite.bgi\_intersected.tsv

* 1. 计算WES TMB

Run：python 3-calculateWesTmb.py

Input: mc3.lite.idt\_intersected.tsv

Output: mc3.lite.tmb.tsv

* 1. 计算Panel TMB

Run：python 3-calculateBgiTmb.py

Input: mc3.lite.bgi\_intersected.tsv； driver.txt

Output: mc3.lite.nodrive.tmb.tsv

* 1. 构建线性回归模型，获取WES TMB（y）和panel TMB（x）之间的线性回归曲线y=a+bx 和决定系数等

Run：Rscript 5-createLinearRegModel.R > EndResult

Input: mc3.lite.nodrive.tmb.tsv

Output: EndResult：

结果中拟合斜率为**1.371559**，截距为**-0.735414**

[1] "Spearman correlation: 0.861074492039215"

[1] "Pearson correlation: 0.992927316810437"

Call:

lm(formula = WES ~ Panel, data = read.table("mc3.lite.nodrive.tmb.tsv",

sep = "\t", header = T, stringsAsFactors = F))

Residuals:

Min 1Q Median 3Q Max

-36.716 -0.743 0.359 1.120 45.248

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.735414 0.040787 -18.03 <2e-16 \*\*\*

Panel 1.371559 0.002178 629.70 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.909 on 5669 degrees of freedom

Multiple R-squared: 0.9859, Adjusted R-squared: 0.9859

F-statistic: 3.965e+05 on 1 and 5669 DF, p-value: < 2.2e-16

**最终我们获得的BGI panel 检测与公共数据库WES 检测的线性回归方程拟合曲线：**

**y=1.371559x-0.735414**

* 1. 获取各癌种的WES 和BGI Panel TMB 的相关性系数（Pearson 和Spearman）  
     Run: python corr.py

Input: mc3.lite.nodrive.tmb.tsv

Output: wes\_panel\_result\_tmb.txt；panel\_result\_tmb.txt ；corr.xls

* 1. 获取各癌种WES 和Panel TMB 的简单线性回归曲线和参数

Run: Rscript 5-createLinearRegModel.R > EndResult

Input：wes\_panel\_result\_tmb.txt；panel\_result\_tmb.txt

Output:

reg.csv（简单线性回归参数（所有癌症））

*示例：*

cancer， 癌症名

values.r.squared， 决定系数

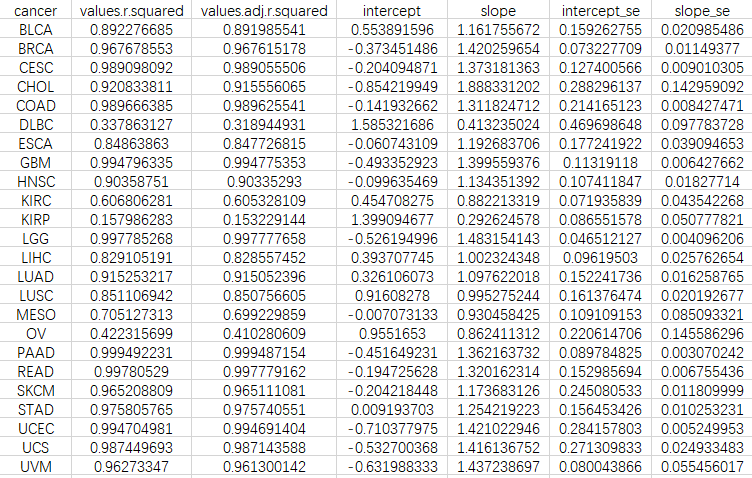
values.adj.r.squared， 调整后决定系数

intercept， 截距

slope， 斜率

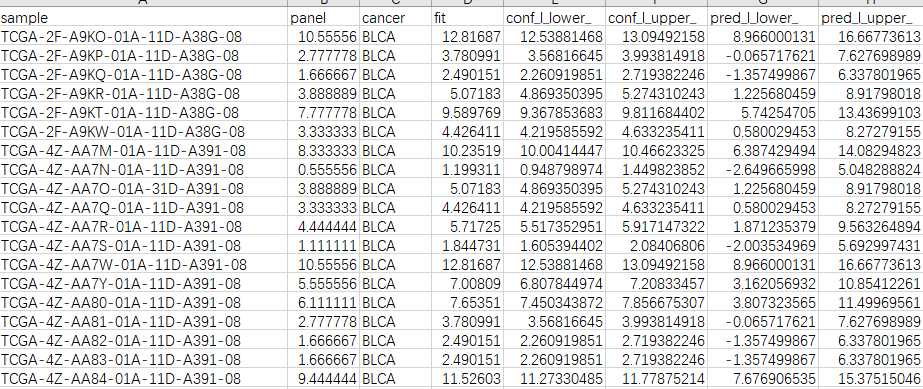
intercept\_se， 截距的standard error

slope\_se， 斜率的standard error



\*.test\_stats.csv（Panel TMB 代入回归式后得到的CI 参数（分癌症））

*示例：*



\*.test\_plot.pdf （Panel TMB 代入回归式后得到的曲线图（分癌症））

*示例：*

