基于对肺鳞癌患者人口特征统计的发现

龚禹桥[1]

[1]上海交通大学生命科学技术学院

摘要：本研究基于最近在 Cell 上发表的一篇题为 A proteogenomic portrait of lung squamous cell carcinoma 的文章，利用其提供的人口特征统计数据，探究了肺鳞癌患者基因突变数和各种因素的关系，构建了相应的判别分析模型，并在进一步的分析中挖掘了基因突变数、相关基因表达量和组织癌变的关系，从而将风险因素、癌相关基因数及表达量和是否癌变联系了起来。

关键字：肺鳞癌、人口统计特征、基因突变数、癌变

引言

肺鳞状细胞癌（又称肺磷癌， lung squamous cell cancer, LSCC）是一种常见肺癌，约占原发性肺癌的 40%～51%，对于肺鳞癌的研究是肺癌研究中的一大热点。近日，国际顶级学术期刊 Cell 上发表题为 “A proteogenomic portrait of lung squamous cell carcinoma” 的文章。[1]研究者通过搜集 108 例未经治疗的原发性肺鳞状细胞癌的肿瘤组织和对应的癌旁组织进行蛋白基因组学分析。其中的数据集为我们提供了 2016 年 5 月至 2018 年 8 月，从 7 个不同国家的 13 个不同组织源部位收集的 108 例肿瘤和 99 对 NATs 的数据，其中详细描述了每个样本来源个体的种族、国家、年龄、性别、吸烟史、癌症分期等信息,且提供了每个样本的 13 个给定基因的突变情况。基于此，我们想从数据集中挖掘出导致患者癌相关基因突变显著升高的风险因素，并进一步探究这些因素与癌变的关系，发现可能的潜在联系，最终可以对普通大众提供可能的健康建议。

方法论和实证分析

风险因素探究

由于给定的数据集中只有 13 个癌相关基因突变的情况，且分布比较离散，更适合作为分类型变量进行处理。因此以 13 个基因的突变数目的中位数 2 作为界限，将 108 个个体分为高突变和低突变（图 1）。其中突变数≥2 的作为高突变组，<1 的作为低突变组。

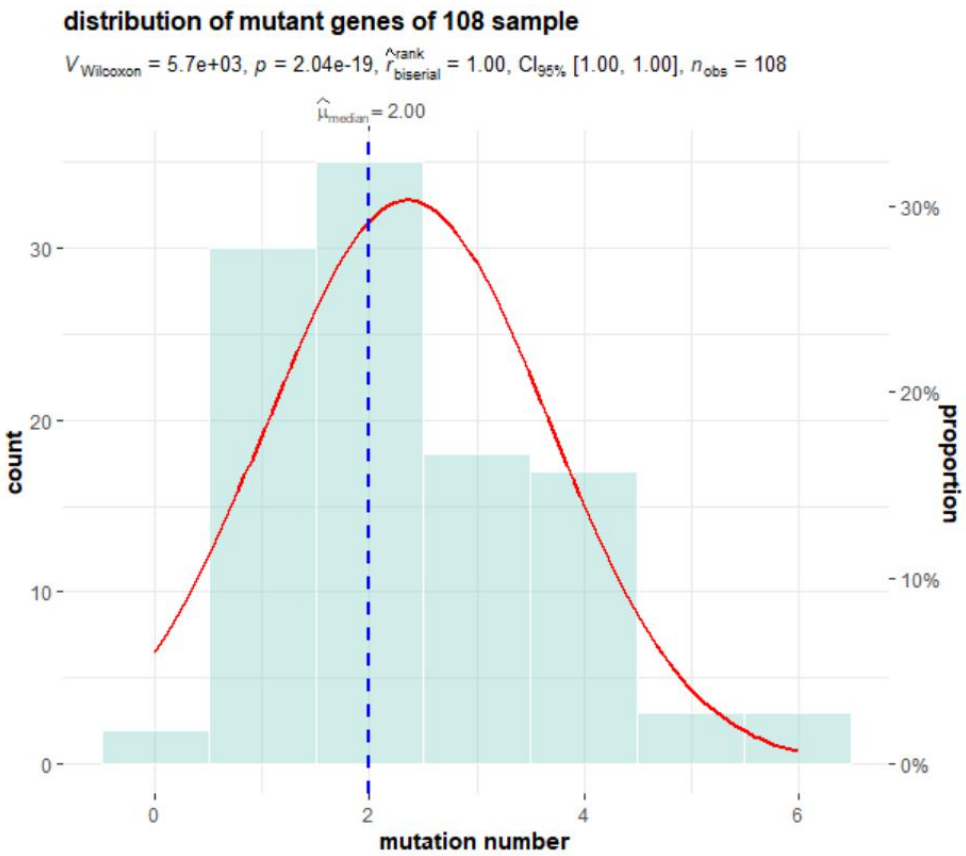
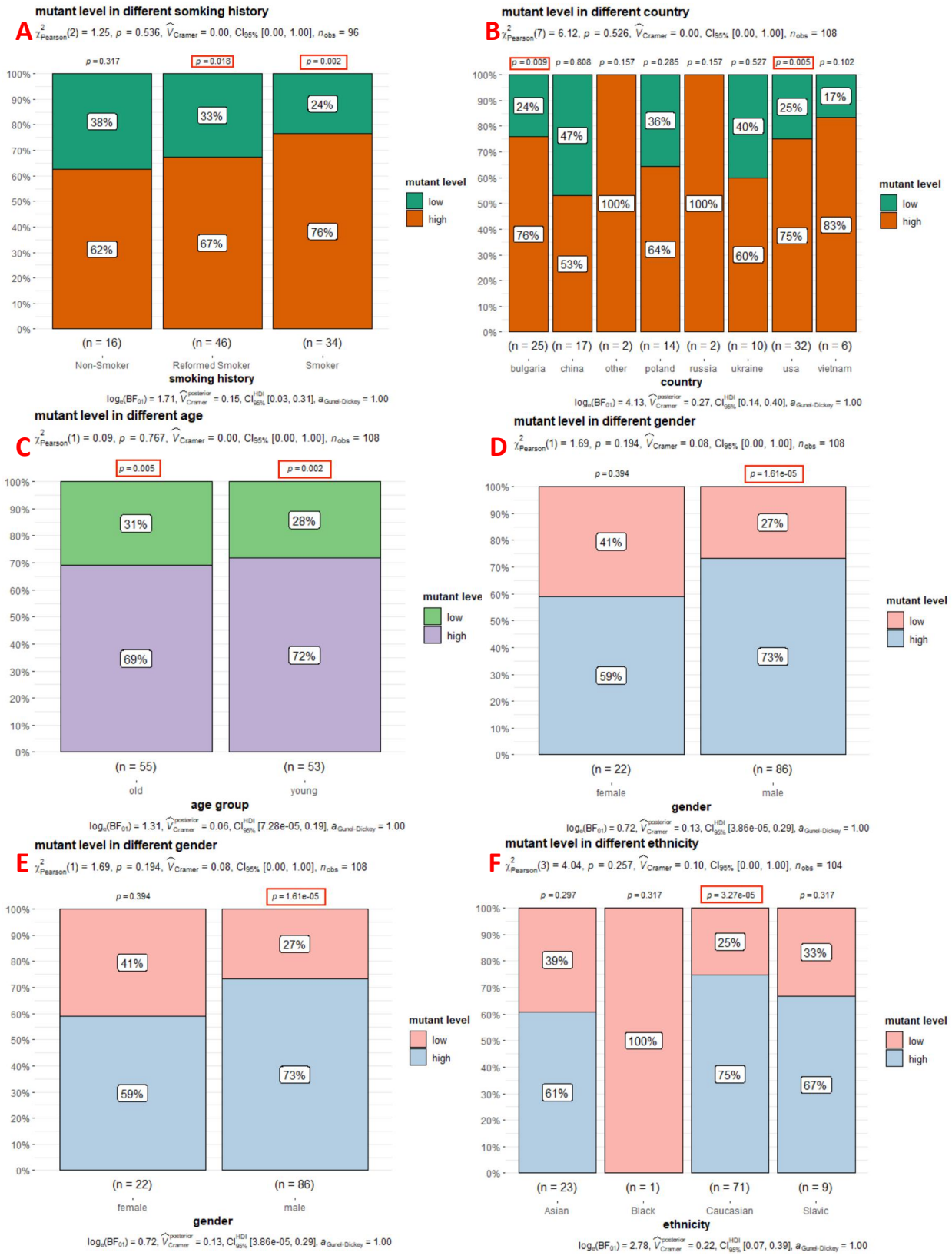


图 1　对于 13 个基因的突变数目进行分组

利用 Chi-Square 检验和单样本比率的检验，检测哪种因素下高突变个体数会显著高于低突变个体数。从结果来看（图 2），虽然在 0.05 的显著性水平下不能得到某一因素下几种水平之间的比例存在显著差异。但值得注意的是，有过吸烟史、美国人或保加利亚人、男性、白种人、每日吸烟数≥20 和吸过二手烟的几种水平下我们得到了高突变个体数显著高于低突变个体数。也就是说这些因素可能会与更高的基因突变数存在关系。

定量回归建模

在得到相关因素的基础上，如果能构建出基因突变数与这些因素的定量回归模型，将会帮助我们更准确地预测一定条件下这 13 个基因的突变数目情况。但是通过散点图观察，这样的数据分布似乎没有明显合适的回归模型。如果尝试建立回归模型,选择基因突变的数目作为响应变量，性别（female=0,male=1）、年龄、日吸烟条数与年吸烟包数为自变量。考虑到响应变量的性质，这里作泊松族的广义线性回归，删除有缺失值的样本，保留了 57 个样本，通过 AIC 准则下的逐步回归后得到的结果只有年龄项达到了显著，但系数-0.02 已经很接近 0，说明在这样的数据上建立回归模型是不合适的。（**图 3**）


A
mutant level in different somking history
$\chi^2_{Pearson}(2) = 1.25$, $p = 0.536$, $\widehat{V}_{Cramer} = 0.00$, $CI_{95\%}$ [0.00, 1.00], $n_{obs} = 96$
p = 0.317
p = 0.018
p = 0.002
38%
62%
33%
67%
24%
76%
(n = 16)
(n = 46)
(n = 34)
Non-Smoker
Reformed Smoker
Smoker
smoking history
mutant level
low
high
$\log_e(BF_{01}) = 1.71$, $\widehat{V}^{posterior}_{Cramer} = 0.15$, $CI^{HDI}_{95\%}$ [0.03, 0.31], $a_{Gunel\text{-}Dickey} = 1.00$
B
mutant level in different country
$\chi^2_{Pearson}(7) = 6.12$, $p = 0.526$, $\widehat{V}_{Cramer} = 0.00$, $CI_{95\%}$ [0.00, 1.00], $n_{obs} = 108$
p = 0.009
p = 0.808
p = 0.157
p = 0.285
p = 0.157
p = 0.527
p = 0.005
p = 0.102
24%
76%
47%
53%
100%
36%
64%
100%
40%
60%
25%
75%
17%
83%
(n = 25)
(n = 17)
(n = 2)
(n = 14)
(n = 2)
(n = 10)
(n = 32)
(n = 6)
bulgaria
china
other
poland
russia
ukraine
usa
vietnam
country
mutant level
low
high
$\log_e(BF_{01}) = 4.13$, $\widehat{V}^{posterior}_{Cramer} = 0.27$, $CI^{HDI}_{95\%}$ [0.14, 0.40], $a_{Gunel\text{-}Dickey} = 1.00$
C
mutant level in different age
$\chi^2_{Pearson}(1) = 0.09$, $p = 0.767$, $\widehat{V}_{Cramer} = 0.00$, $CI_{95\%}$ [0.00, 1.00], $n_{obs} = 108$
p = 0.005
p = 0.002
31%
69%
28%
72%
(n = 55)
(n = 53)
old
young
age group
mutant level
low
high
$\log_e(BF_{01}) = 1.31$, $\widehat{V}^{posterior}_{Cramer} = 0.06$, $CI^{HDI}_{95\%}$ [7.28e-05, 0.19], $a_{Gunel\text{-}Dickey} = 1.00$
D
mutant level in different gender
$\chi^2_{Pearson}(1) = 1.69$, $p = 0.194$, $\widehat{V}_{Cramer} = 0.08$, $CI_{95\%}$ [0.00, 1.00], $n_{obs} = 108$
p = 0.394
p = 1.61e-05
41%
59%
27%
73%
(n = 22)
(n = 86)
female
male
gender
mutant level
low
high
$\log_e(BF_{01}) = 0.72$, $\widehat{V}^{posterior}_{Cramer} = 0.13$, $CI^{HDI}_{95\%}$ [3.86e-05, 0.29], $a_{Gunel\text{-}Dickey} = 1.00$
E
mutant level in different gender
$\chi^2_{Pearson}(1) = 1.69$, $p = 0.194$, $\widehat{V}_{Cramer} = 0.08$, $CI_{95\%}$ [0.00, 1.00], $n_{obs} = 108$
p = 0.394
p = 1.61e-05
41%
59%
27%
73%
(n = 22)
(n = 86)
female
male
gender
mutant level
low
high
$\log_e(BF_{01}) = 0.72$, $\widehat{V}^{posterior}_{Cramer} = 0.13$, $CI^{HDI}_{95\%}$ [3.86e-05, 0.29], $a_{Gunel\text{-}Dickey} = 1.00$
F
mutant level in different ethnicity
$\chi^2_{Pearson}(3) = 4.04$, $p = 0.257$, $\widehat{V}_{Cramer} = 0.10$, $CI_{95\%}$ [0.00, 1.00], $n_{obs} = 104$
p = 0.297
p = 0.317
p = 3.27e-05
p = 0.317
39%
61%
100%
25%
75%
33%
67%
(n = 23)
(n = 1)
(n = 71)
(n = 9)
Asian
Black
Caucasian
Slavic
ethnicity
mutant level
low
high
$\log_e(BF_{01}) = 2.78$, $\widehat{V}^{posterior}_{Cramer} = 0.22$, $CI^{HDI}_{95\%}$ [0.07, 0.39], $a_{Gunel\text{-}Dickey} = 1.00$

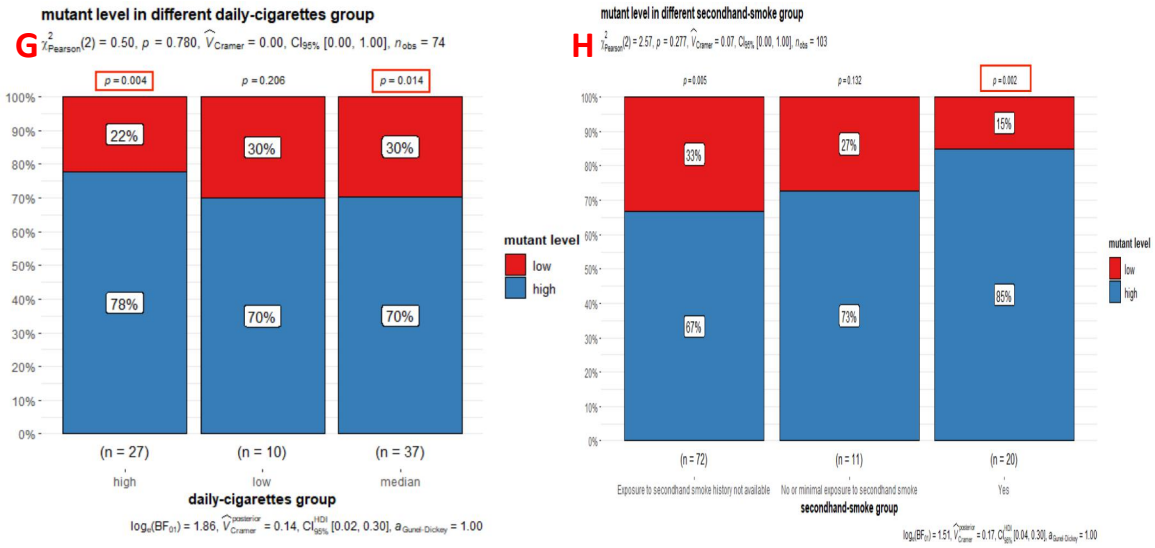


图 2 风险因素探究|A-H 对应的因素分别为吸烟史、国家、年龄、性别、种族、每日吸烟数和是否吸二手烟。其中年龄按照其中位数分为 old 组和 young 组

建立判别模型

由于数据集中大多因素都是分类变量，可能无法用定量的回归模型来描述，但我们可以尝试建立判别模型来对数据进行分类。首先使用距离判别法进行建模，选取年龄、性别、日吸烟条数和吸烟史作为因素。其中吸烟史分为 0,1,2 三个指标，分别对应 non-smoker、reformed-smoker 和 smoker，然而训练组自预测得到的准确率也只有 0.36，说明模型并不好。换用线性判别分析，得到的训练组预测准确率达到了 0.77，测试组准确率也达到了 0.71，说明用线性判别建立的分类模型能够较好地适用于我们的数据。（**图 4**）

挖掘基因突变数和组织癌变的关系

直观的想法是利用数据集给出的癌和癌旁组织基因突变数情况的对比来建立基因突变数和组织癌变的关系。难点在于数据集给出的数据只涵盖了 13 个高变基因的突变情况，而恰好所有的癌旁组织这 13 个基因都没有发生突变。然而正如原文献所言[1],与 LSCC 相关的基因有 700 多个,在不知道其他基因的突变情况下，我们需要找到一条新的线索来联系给定的 13 个基因的突变数和是否发生癌变这一结果。这一思想类似于寻找工具变量（或者中间变量），下面的问题就在于，什么变量作为这个中间变量是比较合适的。以目前的生物学知识，有两个候选的变量可能会既和 13 个基因的突变数有关，又和组织癌变相联系：1.与 LSCC 相关的 700 多个基因的表达量(Total Gene Expression ,TGE)，这样可能会包含更多的潜在信息；2.仅这 13 个基因的的表达量(Significant Gene Expression ,SGE)，这些基因与给定的 13 个基因的突变情况应该会有更直接的关系。

对于第一种情况，我们在转录组数据中挑出原文献列出的 723 个与癌变有关的基因并进行归一化处理,并删除含有缺失值的基因，最终保留了 644 个基因。剔除

掉没有配对的样本，最终保留了 94 对配对的样本。以这些基因的表达量之和作为该样本的总体表达（TGE），考察它和癌变的关系。从结果来看（**图 5A,B**），用 644 个基因的总体表达作为自变量可能太过笼统，其中涉及到的复杂内在联系难以阐述清楚。

第二种情况似乎是更为合适的中间变量（**图 5C,D**）。并且从中得到了一个可能的关系：基因突变数↑⟸Significant Gene Expression (SGE)↑⟺癌变的风险↑

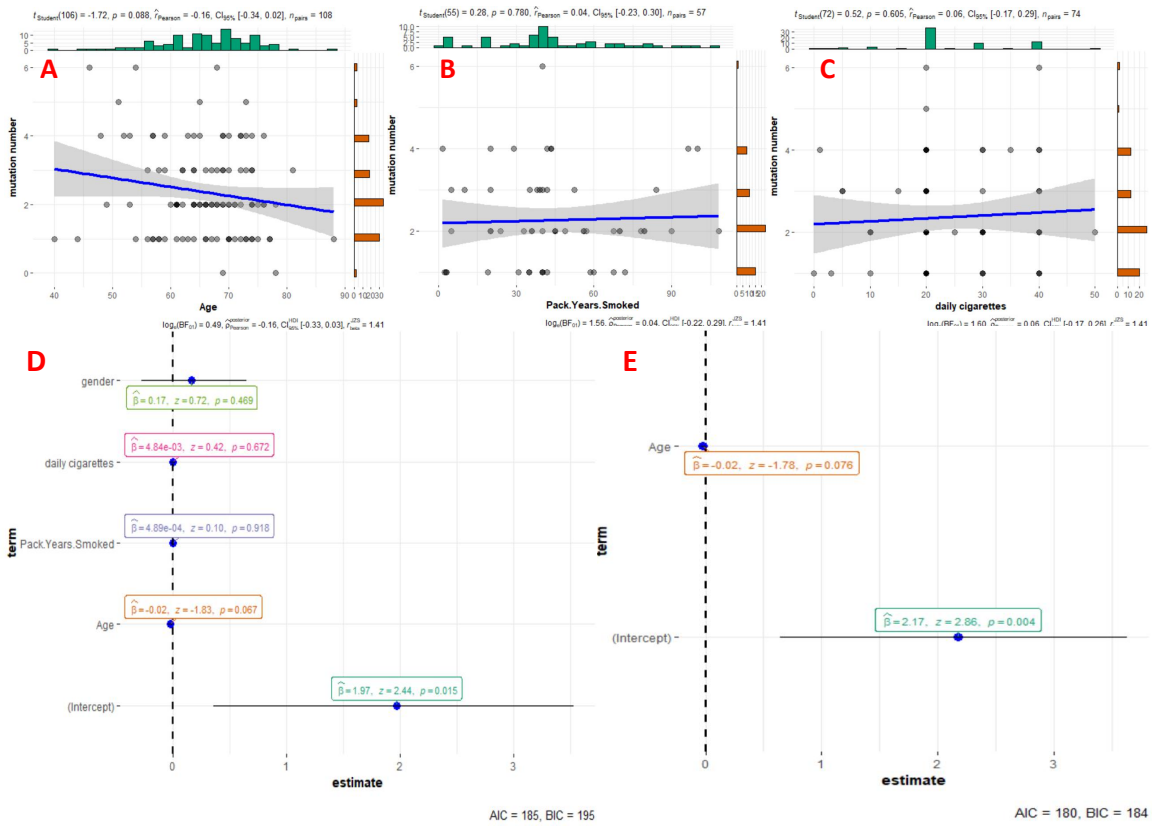


图 3 定量回归建模|A-C 对应的 13 个基因的突变数与年龄、年吸烟包数和日吸烟条数的散点图；D，模型的系数只有截距项显著；E，逐步回归后只剩下 Age 因素显著，且系数接近 0

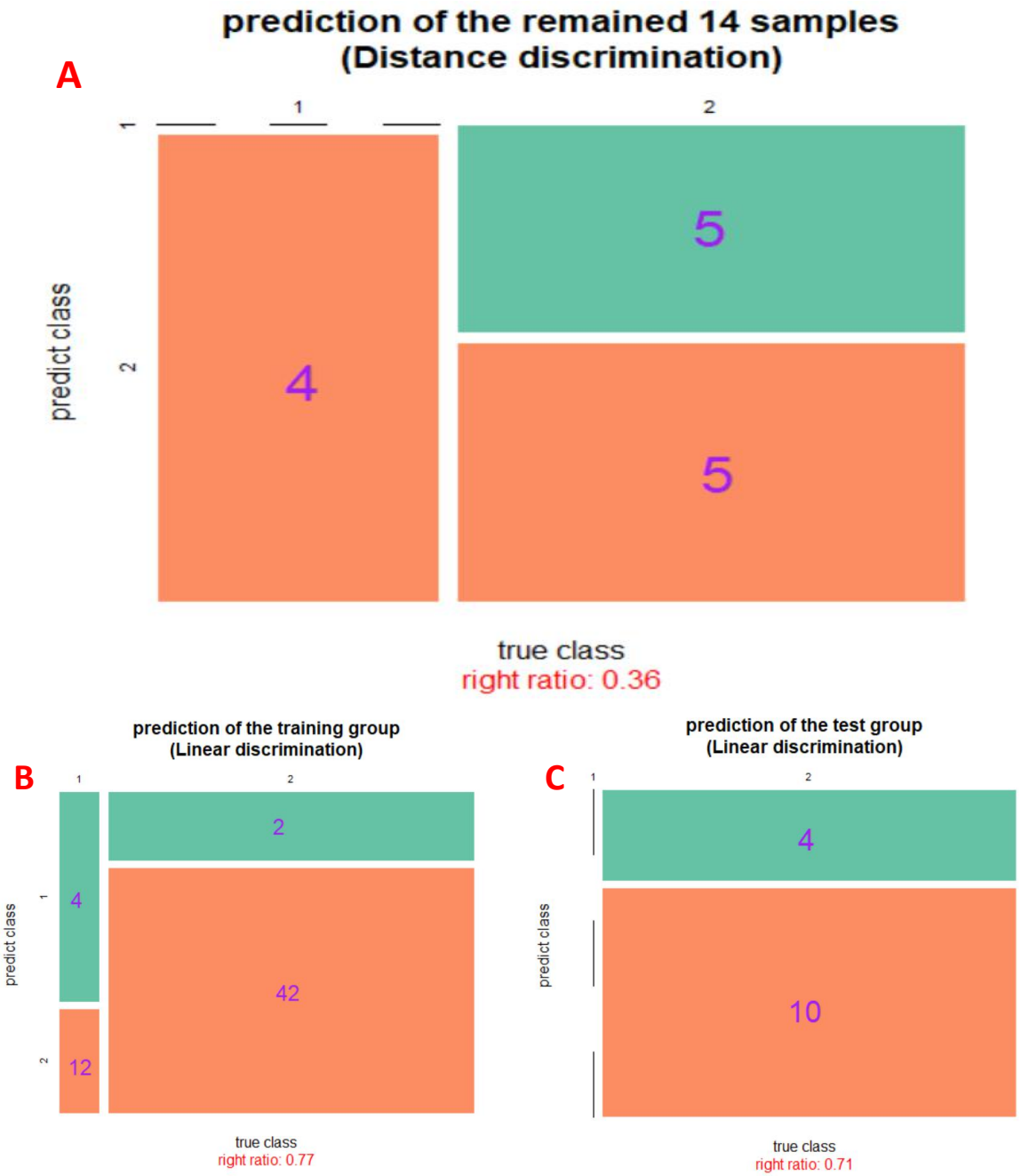


图 4 判别模型预测的混淆矩阵|A，用距离判别法建立的模型准确率很低；B--C，线性判别分析建立的模型在训练集和测试集上的预测效果比较好

揭示潜在的联系

从以上的研究结果，我们可以看到有过吸烟史、美国人或保加利亚人、男性、白种人、每日吸烟数≥20 和吸过二手烟的几种水平下在给定的 13 个基因中高突变个体数显著高于低突变个体数。这些因素可能会与更高的基因突变数存在关系。而我们又通过研究发现更高的基因突变数往往对应更高的癌变的风险。这提示我们发现的这些因素可能会与更高的患癌风险相联系。也为我们指出了可能的健康方面的建议。

总结和展望

在本研究中，我们以探究不同因素下高低水平变异的人数区别为出发点，揭示了有过吸烟史、美国人或保加利亚人、男性、白种人、每日吸烟数≥20 和吸过二手烟的几种水平下在给定的 13 个基因中高突变个体数显著高于低突变个体数，从而得到这些因素与基因的高突变之间的联系。之后又通过寻找中间变量 SGE 的方式在数据有限的情况下挖掘出了 13 个基因的突变数目与癌变之间的关系，进而揭示了几种因素和患癌风险之间的潜在联系。这一研究为可能的健康建议提出了指导。

可能的不足之处在于设计高低水平变异的分组时，以中位数来划分可能不是一个好的方式。另外由于正好处于中位数的个体数过多，导致这一部分的个体无论划分到哪个组都会对结果有比较大的影响。可能这也是在多组卡方检验时不显著的主要原因。可能的改进的方案是将这部分个体单独作为一组，将所有个体分为三组来进行实验。

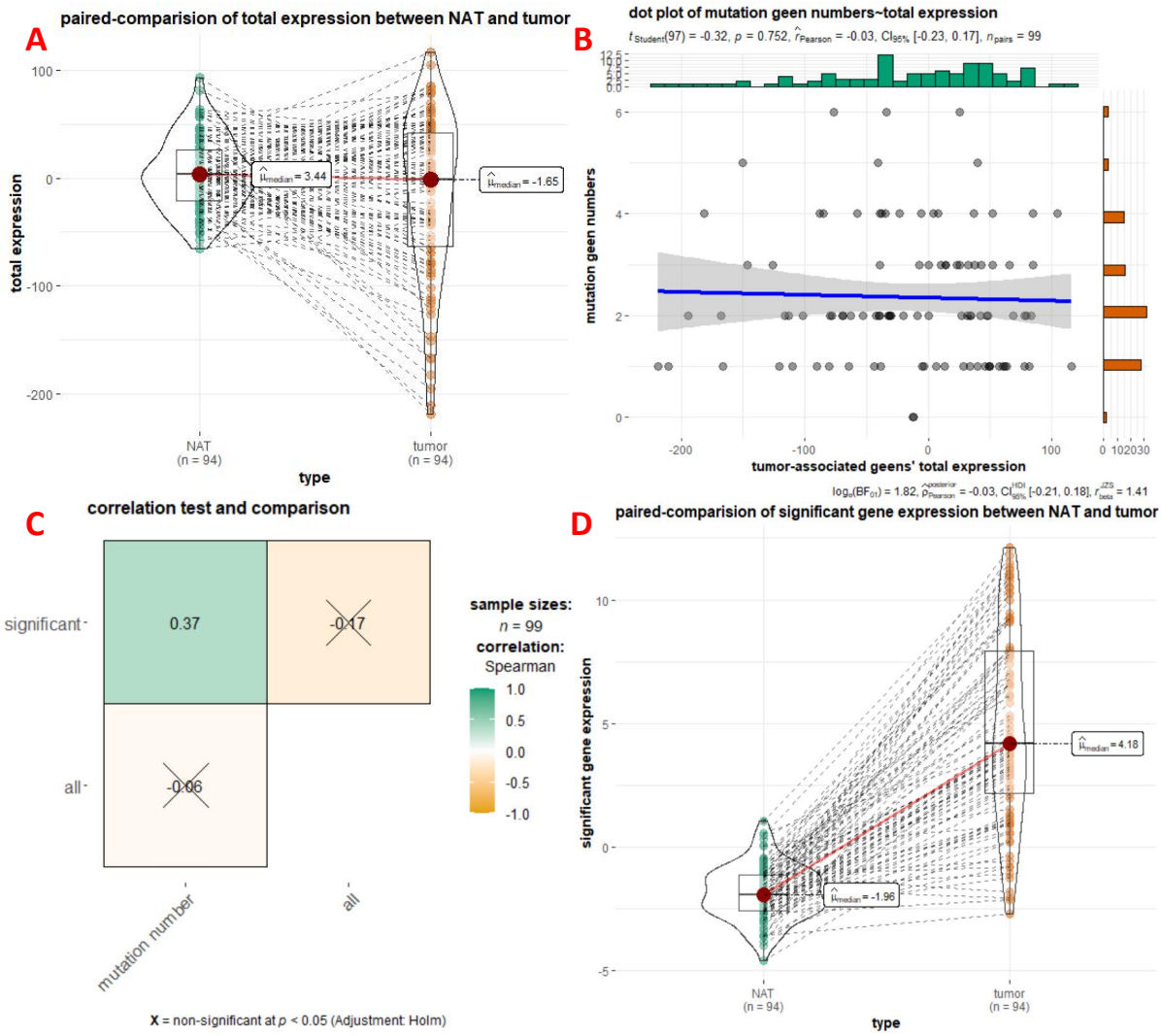


图 4 寻找中间变量|A，癌和癌旁 TGE 的配对 wilcoxon 检验；B，TGE 和 13 个基因突变数目的关系；C，SGE 和 13 个基因突变数目的相关性检验；D，癌和癌旁 SGE 的配对 wilcoxon 检验

附录

原文献数据集：

https://www.sciencedirect.com/science/article/pii/S0092867421008576?via%3Dihub

分析流程代码：

https://github.com/GYQ-form/Cell-analyse