# 基于肺癌全基因组关联研究数据的疾病风险预测[1]

南京医科大学　　段巍巍、张秋伊、陈海

## 摘要

**目的** 探讨基于肺癌全基因组关联研究数据的风险预测效果。**对象和方法** 本研究数据来源于本校公共卫生学院分子流行病学实验室完成的中国汉族人群非小细胞肺癌 GWAS 研究。将该数据按一定比例随机分成训练集和测试集；在训练集中构建三个预测模型--仅包含传统因素（模型 1）、传统因素加 sGRS 位点集（模型 2）、传统因素加 sGRS-LMM 位点集（模型 3），并考察参数的最优取值组合（假设检验初筛水准和连锁不平衡修剪参数）；最后在测试集中比较三个模型下的预测准确度。**结果** 在指定初筛水准和修剪参数下, 测试集中预测模型 1、2、3 的 AUC 值分别为 0.687（95%可信区间：0.656~0.719）、0.831（95%可信区间：0.807~0.854）、0.847（95%可信区间：0.824~0.869），模型 2、3 与模型 1AUC 比较的假设检验 *P* 值分别为 1.5E-17 和 1.7E-21。**结论** 相较于传统因素，常见变异位点可以较大幅度地提高肺癌的预测准确度，sGRS 和 sGRS-LMM 两种方法可以用于肺癌全基因组关联研究数据的疾病风险预测。

**关键词** 肺癌；风险预测；单核苷酸多态性；全基因组关联研究

[1] 注:该论文获得由中国统计教育学会举办的“ 2015 年（第四届）全国大学生统计建模大赛 ”大数据统计建模类研究生组一等奖。

传统流行病学研究已经发现了大量的疾病相关风险因素，而基于这些因素的疾病风险预测也取得了一定的应用[1]。随着人类遗传学研究技术的发展，遗传因素对疾病发生的影响引起了人们的足够重视。近年来，全基因组关联研究（genome-wide association study，GWAS）已经被认为是阐明复杂性状遗传关联机制的强有力工具。截止 2015 年 02 月底，全球的研究者们累计发现了与 1,251 种性状（疾病）相关的 19,602 个单核苷酸多态性（Single Nucleotide Polymorphism，SNP）位点[2]。充分利用这些已发现的位点并结合传统风险因素进行临床个体化医疗实践、疾病预防等应用成为后 GWAS（post-GWAS）时代的主要目标之一，而这最为关键的一步则在于建立一个准确的风险预测模型。早期的风险预测研究表明，利用 GWAS 获得的关联位点进行预测并不理想[3-5]，其中的一个主要原因在于这些研究忽略了大量存在、未被发掘的低效应位点，因此如何充分利用 GWAS 研究信息成为预测模型成败的关键。近来，主要有两类策略被提出，表现出较好的预测效果：一类是通过设定宽松的假设检验水准[6, 7]，以便纳入一些潜在的关联位点；另一类是借助于混合效应模型将全基因组所有常见变异位点纳入预测模型[8, 9]。本研究基于上述策略提出两种改进方法和一个多步骤的风险预测分析流程，考察相对于传统因素，基于 SNP 位点的遗传因素能增加预测效果的程度，为 GWAS 数据的风险预测提供参考。

## 一、材料与方法

### （一）研究对象

本研究数据来源于南京医科大学公共卫生学院分子流行病学实验室完成的中国汉族人群非小细胞肺癌 GWAS 研究[10]，包含两个子研究：南京子研究和北京子研究。所有病例都经过细胞学或组织病理学的确证，对照则是从当地医院的体检人员中选取或者是从非传染性疾病的筛检项目中获得。对照组与病例组在性别构成上均衡可比。所有的研究对象均有完整的流行病学调查资料，包括一般情况、家族史、个人暴露史、疾病史、临床资料(包括诊断和治疗)等。每天吸烟少于一支或者一生吸烟少于一年的人定义为非吸烟者，吸烟超过一年但已经停吸的人定义为曾吸烟者，癌症家族史的定义是任何一级亲属(包括父母、兄弟姐妹、子女)的自报癌症情况。此项研究通过了每个研究机构审查委员会的审查。

### （二）基因分型与质量控制

所有研究个体的 DNA 样本来自于全血，使用 Affymetrix Genome-Wide Human SNP Array 6.0 芯片进行基因分型，分型的具体方法步骤在之前的研究已

有详细描述[10]。按如下质量控制要求删除位点：①SNP 位点位于 22 条常染色体之外；②总样本及研究子样本（南京和北京子研究）缺失率大于 5%；③MAF 小于 0.05；④总样本 Hardy-Weinberg 平衡检验 *P* 值小于 $10^{-5}$，子样本检验 *P* 值小于 $10^{-4}$；最终，共计剩余 570,373 个位点。按如下要求剔除研究个体：①性染色体的遗传性别与问卷性别不一致；②基因分型成功率小于 95%；③存在杂合性缺失；④存在亲缘关系；⑤人群分层中的离群（outlier）个体。经过严格的质量控制后, 共计 570,373 个 SNP 位点用于分析, 5408 个研究样本（2331 个病例和 3077 个对照），其中两个子样本：南京样本（1473 个病例和 1962 个对照）和北京样本（858 个病例和 1115 个对照）。

## （三）SNP 遗传度估计

此处简单介绍 GWAS 研究中遗传度与遗传方差的估计方法，以及遗传度与预测指标之间的关系，以此来指导和评价实例数据中的风险预测研究。后面提及的遗传预测方法也将基于此部分内容。

1. 遗传度

遗传度(heritability)，又称遗传力，指的是群体在某一表型上所表现出来的差异归因于遗传的比例。对于表型（phenotype, *P*），可以将之分解为基因型（genotype, *G*）贡献的部分和环境因素（environment, *E*）贡献的部分，即：*P*=*G*+*E*。当 *G* 与 *E* 独立时，则有表型方差：

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 \quad (1)$$

其中, $\sigma_G^2$为遗传方差, $\sigma_E^2$为环境方差, 则有广义（board sense）遗传度$H^2 = \sigma_G^2/\sigma_P^2$。如果把遗传方差继续分解：

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 \quad (2)$$

其中，$\sigma_A^2$为表示育种值（breeding value）的加性遗传方差，为可以固定遗传的变异，$\sigma_D^2$为表示同一位点等位基因间交互作用的显性（dominance）遗传方差，$\sigma_I^2$为表示不同位点等位基因间交互作用的上位性（epistatic）方差[11], 则有狭义（narrow sense）遗传度$h^2 = \sigma_A^2/\sigma_P^2$。

2. 遗传方差与遗传度估计

狭义遗传度（以下简称“遗传度”）的估计方法有多种，一般针对不同的研究设计与数据而采用对应的方法。对于简单的平衡设计，可以通过对家系研究中

父母表型与子代表型做线性回归获得遗传度估计值；对于存在多种亲属关系混合的研究对象或采用非平衡设计时，则通常采用随机效应模型。Yang 等人[12, 13]首次将传统的遗传度估计方法应用于 GWAS 研究的高密度分型数据，并解释了所谓的“丢失的遗传度”（missing heritability）[14]问题（即 GWAS 发现的少量位点仅能解释一小部分的遗传方差），他们认为：GWAS 研究中存在大量的低效应位点，而这些位点并不能达到关联阈值；与此同时，部分或全部的真实关联位点与标签 SNP 之间的连锁不平衡（linkage disequilibrium, LD）并不完美。通过这些 SNP 位点所估计出的遗传度，一般称之为芯片遗传度（chip-heritability）或 SNP 遗传度（SNP-heritability），以下将简称遗传度。对于一般的线性混合效应模型：

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{u} + \boldsymbol{e} \tag{3}$$

其中，$\boldsymbol{Y}$ 为 $N\times1$ 的表型向量，$N$ 为样本含量，$\mathbf{X}$ 为协变量矩阵，$\boldsymbol{\beta}$ 为对应协变量的效应值，$\mathbf{G}$ 为标准化后的 SNP 基因型矩阵，$\boldsymbol{u}$ 为 SNP 效应向量，假定其满足 $\boldsymbol{u}\sim N(0, \mathbf{I}\sigma_u^2)$，$\mathbf{I}$ 为单位阵，误差项 $\boldsymbol{e}\sim N(0, \mathbf{I}\sigma_e^2)$，则 $\boldsymbol{Y}$ 的方差为 $var(\boldsymbol{Y}) = \mathbf{G}\mathbf{G}'\sigma_u^2 + \mathbf{I}\sigma_e^2$，令 $\mathbf{A} = \mathbf{G}\mathbf{G}'/n$ 和 $\sigma_g^2 = n\sigma_u^2$，其中 $n$ 为 SNP 位点数，此时 $\sigma_g^2$ 即为全部 SNP 位点所解释的遗传方差，矩阵 $\mathbf{A}$ 是遗传相似矩阵（genetic relationship matrix, GRM），$\sigma_g^2/(\sigma_g^2 + \sigma_e^2)$ 即为观察尺度（observed scale）下估计的遗传度 $h_o^2$。考虑到对方差成份（variance component）的无偏估计，求解该模型通常采用限制性极大似然估计（restricted maximum likelihood, REML），为了加快迭代过程的收敛，通常采用平均信息（average of observed and expected information, AI）算法[15, 16]，该混合效应模型的对数似然方程[17]为：

$$L = -1/2(ln|\mathbf{V}| + ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \boldsymbol{Y}'\mathbf{P}\boldsymbol{Y}) \tag{4}$$

此处，投影矩阵 $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$，$\mathbf{V}$ 为表型的方差-协方差矩阵。基于 AI 的 REML 迭代过程为：

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + (\mathbf{I}_{\mathrm{A}}^{(k)})^{-1}\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}^{(k)}} \tag{5}$$

其中，$\boldsymbol{\theta}$ 为各方差成分组成的向量，对于本研究 $\boldsymbol{\theta} = (\sigma_g^2, \sigma_e^2)'$，$k$ 为迭代的次数，$\mathbf{I}_{\mathbf{A}}$ 为平均信息矩阵，它包含了对数似然函数 $L$ 关于方差成分的二阶导数，在本研究则有：

$$\mathbf{I}_{\mathbf{A}} = \frac{1}{2}\begin{bmatrix} Y'\mathrm{PAPAP}Y & Y'\mathrm{PAPP}Y \\ Y'\mathrm{PPAP}Y & Y'\mathrm{PPP}Y \end{bmatrix}$$

而$\partial L/\partial \boldsymbol{\theta}$则是关于方差成分的一阶导数：

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -\frac{1}{2}\begin{bmatrix} \mathrm{tr(PA)} - Y'\mathrm{PAP}Y \\ \mathrm{tr(P)} - Y'\mathrm{PP}Y \end{bmatrix}$$

该迭代的收敛准则为相邻两次迭代的对数似然之差小于 $10^{-4}$, 即 $\boldsymbol{L}^{(k+1)}$-$\boldsymbol{L}^{(K)}$<$10^{-4}$。其它进行 REML 的方法还包括牛顿迭代（Newton-Raphson method）[18]、Fisher 得分[19]和谱分解（spectral decomposition）等。

对于二分类表型，估计的方法基本相同，将该二分类表型当作连续性表型替代公式（3）中的 $Y$，假设此时被估计出的遗传度为$h_o^2$，考虑到该值是人群参数的函数：对于人群患病率为 $K$ 的某一疾病，其表型方差为 $K$(1-$K$)。由于均数与方差存在函数关系, 因此遗传因子的重要性往往随着人群患病率的不同而变得不同。为了获得一个不受人群参数影响、可以与其它疾病进行直接比较的遗传度，通常会使用倾向阈值模型（liability threshold model）[20]对上述$h_o^2$进行转换，即认为在该模型下有一个潜在的连续随机变量，大于某一个阈值时则认为发生疾病，而小于阈值则认为不发生，此连续变量对应的遗传度即为潜在的遗传度$h_l^2$（liability heritability)。如果该随机变量满足标准正态分布时，可以用图 1 来描述该模型，阈值左右曲线下围成的面积分别为 1-$K$ 和 $K$。$h_l^2$与$h_o^2$存在如下关系[21, 22]：

$$h_l^2 = h_o^2 K(1-K)/z^2 \tag{6}$$

而病例-对照研究的样本并非总体人群的随机样本, 此时两者关系做如下修正[23]：

$$h_l^2 = \frac{h_o^2 K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)} \tag{7}$$

此处，$z$ 为标准正态分布下阈值对应的概率密度，$P$ 为病例-对照研究病例样本量所占全部样本量的比例。$h_l^2$的方差估计可以通过$h_o^2$的方差转化而来，不过此时需借助泰勒展开式。转换后的遗传度独立于疾病参数，因此可以进行不同人群、不同性状的比较。

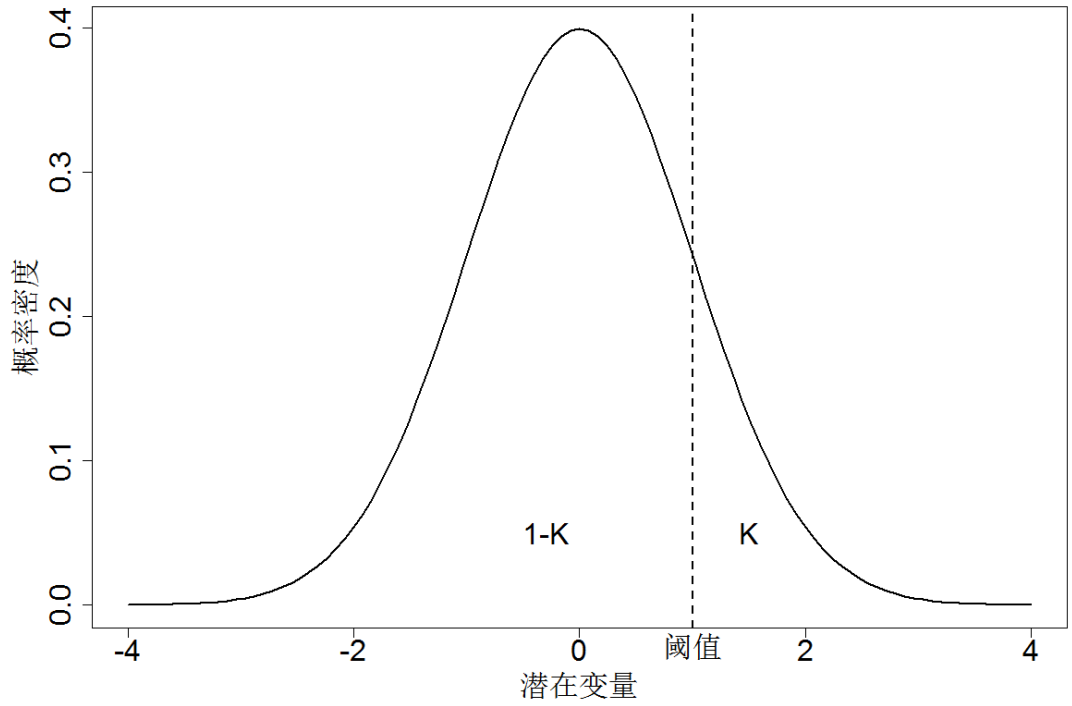


图 1　患病率为 K 的倾向阈值模型

3. 遗传度与 AUC

在遗传方差估计的过程中，令总体遗传值 $\boldsymbol{g}=\mathbf{G}\boldsymbol{u}$，$\hat{\boldsymbol{g}}$为 $\boldsymbol{g}$ 的估计值，则真实值与估计值的相关系数$\rho_{\boldsymbol{g}\hat{\boldsymbol{g}}}=\frac{Cov(\boldsymbol{g},\hat{\boldsymbol{g}})}{\sqrt{Var(\boldsymbol{g})Var(\hat{\boldsymbol{g}})}}$，由于$\hat{\boldsymbol{g}}$是遗传值 $\boldsymbol{g}$ 的无偏估计，则有 $Cov(\boldsymbol{g},\hat{\boldsymbol{g}})=Var(\hat{\boldsymbol{g}})$，即$\rho_{\boldsymbol{g}\hat{\boldsymbol{g}}}=\sqrt{Var(\hat{\boldsymbol{g}})/Var(\boldsymbol{g})}$。遗传度与 AUC 的关系比较复杂，它们受到了疾病的遗传流行病学因素（如疾病患病率）影响。本研究直接利用 Naomi R.W 等人[24]的近似结果：

$$\mathrm{AUC}_{\max}\approx\Phi\left(\frac{(i-v)h_l^2}{\sqrt{h_l^2[(1-h_l^2 i(i-T))+(1-h_l^2 v(v-T))]}}\right) \tag{8}$$

其中，$\Phi(\cdot)$为标准正态分布的累积概率密度函数，$i=z/K$，$z$ 与 $K$ 的含义参见前一节，$v=-iK/(1-K)$，$T=\Phi^{-1}(1-K)$，$h_l^2$为潜在的遗传度。$\mathrm{AUC}_{\max}$ 的含义在于，当$\rho_{\boldsymbol{g}\hat{\boldsymbol{g}}}^2=1$（即遗传值方差的估计值等于其真实值）时，遗传风险预测研究中 AUC 能取得理论值。由于 $\mathrm{AUC}_{\max}$ 在现实情况下较难达到，本研究也同时计算 $\mathrm{AUC}_{\mathrm{half}}$，即当$\rho_{\boldsymbol{g}\hat{\boldsymbol{g}}}^2=0.5$ 时的理论 AUC 值，计算时只需要将上述公式中的$h_l^2$替换为 $0.5h_l^2$即可。

## （四）研究策略与统计分析

1. 风险预测评价指标

风险预测的评价指标有多个，如受试者工作特征曲线（receiver operating characteristic curve，ROC）下面积（area under the curve, AUC）、再分类优值（net reclassification improvement，NRI）和综合判别优值（integrated discrimination improvement, IDI）等，本研究将选取 AUC 作为风险预测评价指标。求解 AUC 的算法有多种，由于本研究数据计算量巨大，为了提高计算速度，可以近似通过 Mann-Whitney 检验统计量[25, 26]来估计 AUC 值：

$$\widehat{\mathrm{AUC}} = U_{\mathrm{c}}/n_{case}n_{control} \tag{9}$$

其中，$U_{\mathrm{c}}$ 为检验统计量，$n_{case}$ 和 $n_{control}$ 分别表示病例组和对照组的样本含量。AUC 理论取值范围为 0.5~1，取值越大，表明预测效果越好。AUC 的可信区间以及两个具有相关关系 AUC 比较可以使用 Bootstrap 方法或 DeLong[27, 28]等人提出的非参数方法，而本研究中使用了 DeLong 的方法。

2. 逐步加权遗传得分和混合效应模型

逐步加权遗传得分（stepwise weighted GRS, sGRS）是本研究在遗传风险得分（genetic risk score, GRS）的基础上做了改进，将能否最大提升预测效果作为纳入某一候选位点的标准。方法如下：首先在训练集中进行位点初筛，即每一个位点都与表型变量做 logistic 或线性回归。随后将仅对假设检验 $P$ 值小于初筛水准 $\alpha$ 的位点进行分析：

（1）对于所有初筛后的 $m$ 个位点，选择其中具有最佳预测准确度的位点。对于连续表型，有标准化后基因型矩阵 $G$ 和表型向量 $y$，令相关系数向量 $Corr=Gy*\mathrm{sign}(\hat{\beta})$，则向量中最大值对应的位点即为 $m$ 个位点中的最佳预测位点，此处$\hat{\beta}$为每个位点回归系数组成的向量；同理，对于二分类表型，则选择 $m$ 个位点中受试者工作曲线下面积最大的位点。

（2）然后从剩下的 $m$-1 个位点中挑选一个位点纳入预测模型，使得新模型的预测能力提升最大。当满足$argmax_{\mathrm{i}}\{Performance[\mathrm{Phenotype}, G\hat{\beta} + G_i\hat{\beta}_\iota]\}$时，则选入第 i 个位点，其中，$G\hat{\beta}$表示现有模型的遗传得分，$G_i\hat{\beta}_\iota$表示待加入模型的位点的遗传得分。

（3）重复过程（2），直到纳入某一个位点后模型的预测能力不再提高时，

则停止搜索。此时纳入模型的 SNP 位点即组成了预测子集。

而近些年来，因为混合效应模型（linear mixed models, LMMs）较好地控制人群混杂的良好性质[29]，已经被大量应用到全基因组关联研究及其后续研究中。为了进一步纳入 sGRS 中丢掉的位点，本研究将剩余位点以随机效应的方式放入预测模型中，提出了基于 sGRS 的混合效应模型（sGRS-based linear mixed model, sGRS-LMM）的方法。

本研究将 sGRS 获取的得分向量 $\boldsymbol{S}$ 纳入混合效应模型中，即在公式（3）的基础上添加固定效应部分 $\boldsymbol{S}$：

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{S}\beta_1 + \mathbf{G}\boldsymbol{u} + \boldsymbol{e} \tag{10}$$

此处，$\boldsymbol{S}$ 为 sGRS 得分向量，$\beta_1$为其效应，此时的$\boldsymbol{G}$为剩余位点组成的基因型矩阵，其他矩阵或向量符号含义同前。通过 REML 获得表型的方差-协方差矩阵 $\mathbf{V}$ 的估计值$\hat{\mathbf{V}}$，然后借助于混合模型方程（mixed model equation）求解固定部分（协变量和 sGRS 得分变量）效应：

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\beta}_1 \end{pmatrix} = \left( \begin{bmatrix} \boldsymbol{X} & \boldsymbol{S} \end{bmatrix}' \hat{\boldsymbol{V}}^{-1} \begin{bmatrix} \boldsymbol{X} & \boldsymbol{S} \end{bmatrix} \right)^{-1} \begin{bmatrix} \boldsymbol{X} & \boldsymbol{S} \end{bmatrix}' \hat{\boldsymbol{V}}^{-1} \boldsymbol{Y} \tag{11}$$

和随机部分效应 $\boldsymbol{u}$：

$$\hat{\boldsymbol{u}} = \hat{\sigma}_u^2 \boldsymbol{G}' \hat{\boldsymbol{V}}^{-1} \left( \boldsymbol{Y} - \begin{bmatrix} \boldsymbol{X} & \boldsymbol{S} \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\beta}_1 \end{pmatrix} \right) \tag{12}$$

其中$\hat{\boldsymbol{\beta}}$与$\widehat{\beta_1}$分别为$\boldsymbol{\beta}$、$\beta_1$的最佳线性无偏估计（best linear unbiased estimation, BLUE），$\hat{\boldsymbol{u}}$为 $\boldsymbol{u}$ 的最佳线性无偏预测（best linear unbiased prediction, BLUP）[30]，$\widehat{\sigma_u^2}$为 REML 方法求得的$\sigma_u^2$估计值。令遗传值向量 $\boldsymbol{g}$=$\mathbf{G}\boldsymbol{u}$，则$\hat{\boldsymbol{g}}$为 $\boldsymbol{g}$ 的 BLUP：

$$\hat{\boldsymbol{g}} = \mathbf{G}\hat{\boldsymbol{u}}$$

如果要在独立的外部数据集中预测其遗传值，则需要把上式中的 $\mathbf{G}$ 矩阵换成外部数据集对应的基因型矩阵即可。获取遗传值$\hat{\boldsymbol{g}}$后，即可以与相关表型做 ROC 曲线。

对于肺癌病例-对照研究的二分类表型数据，考虑到高维数据下求解广义线性混合效应模型（generalized linear mixed models, GLMMs）的算法复杂度，此处参照遗传方差估计的方法，直接将二分类表型当作连续性变量放入模型，而不需

要做任何的变换。相对于 sGRS，结合混合效应模型的方法将未被 sGRS 包含的可能低效应位点当做随机项纳入混合效应模型, 充分利用了全部的常见变异位点。

3. 研究策略

从全部研究对象中随机抽取了 1081 个（20%）研究对象作为外部验证数据，对剩余研究人群（B）进行五倍交叉验证（5-fold cross-validation），来筛选最佳模型参数。研究步骤如下：

（1）单位点 logistic 回归：将该 GWAS 分型的 57 万个 SNP 位点分别与表型变量做 logistic 回归，与此同时校正外部协变量，包括年龄、性别、吸烟因素和人群分层前十个主成分。其中，前三个变量通过问卷调查获得，人群分层主成分则通过对矩阵$\mathbf{A}=\mathbf{GG}'/\mathrm{n}$进行谱分解，即$\mathbf{A}=\mathbf{Q\Lambda Q}'$，$\mathbf{\Lambda}$为特征值组成的对角矩阵，$\mathbf{Q}$ 的每一列为对应的特征向量，即主成分。位点分型数据采用相加生物学模式（additive model）：假设某一位点有 A、B 两个等位基因，其中 A 为参考等位基因（reference allele），则数据编码为 0、1 或 2，分别对应 AA、AB 和 BB。

（2）SNP 位点初筛：为了考虑初筛水准（$\alpha$）和连锁不平衡（LD）结构对预测准确度的影响，本研究设定多个 $\alpha$ 和基于 LD 结构进行 SNP 位点修剪（LD-Pruning）阈值，此处取 $\alpha$ 的负对数值，即-$\log_{10}(\alpha)$取值为 2~7，LD-Pruning 参数为 $r^2$<1、$r^2$<0.5 和 $r^2$<0.2，分别表示不修剪任何 SNP 位点、修剪掉 $r^2$ 大于 0.5 的位点、$r^2$ 大于 0.2 的位点。筛选后的位点进行下一步分析。

（3）构建预测模型：基于 sGRS 和 sGRS-LMM 两种方法构建风险预测模型。通过比较不同 $\alpha$ 和 $r^2$，以交叉验证下测试集 AUC 的均数作为获取最佳组合的标准。

（4）外部验证：将 B 人群作为训练集，根据步骤（3）选择的最佳 $\alpha$ 和 $r^2$ 组合构建模型，对事先确定的外部验证数据进行预测，并进行三类模型的预测效果比较：①仅包含传统预测因素；②传统因素+sGRS 位点；③传统因素+sGRS-LMM 位点。

与此同时，可以对肺癌芯片遗传度进行估计，并根据公式（8）计算出理论 AUC 值，从而比较理论值和实际值的差距，指导后续遗传风险预测研究。

4. 软件使用

分型位点和研究样本的质量控制、数据处理、关联研究（logistic 回归）以及遗传得分的计算均使用 plink v1.07[31]，遗传度与遗传方差的估计、主成分提取借助于 GCTA v1.24.4[13]，其他计算及处理均使用 R 软件来完成。

## 二、研究结果

### （一）遗传方差分析

本研究对该肺癌 GWAS 数据进行遗传方差分析，结果见表 1。同样，为了考虑人群结构对遗传度估计的影响，此处亦校正了主成分。根据公式（7），观察到的遗传度需要经过倾向阈值模型转换，此时需要人群患病率参数，由于难以获得相应人群非小细胞肺癌的患病率，此处将使用全国肺癌人群流行病调查数据作近似替代：该 GWAS 研究对象为 2003 年至 2011 年医院收集的新发非小细胞肺癌病例，故此处使用国家癌症中心公布的历年肺癌发病率[32-34]作为此处人群患病率的估计值，约为 $40/1\times10^5$。此外，根据公式（8）估计出遗传度对应的理论 AUC 值（$AUC_{max}$）和达到一半遗传值变异的理论 AUC 值（$AUC_{half}$）。

表 1　基于 GWAS 常见变异位点的遗传度分析

| 人群 | 样本量[a]（病例/对照） | [b]$h^2$ | 标准误 SE($h^2$) | $AUC_{max}$ | $AUC_{half}$ |
|---|---|---|---|---|---|
| 总人群 | 5408（0.431） | 0.271 | 0.017 | 0.923 | 0.835 |
| 南京子人群 | 3435（0.429） | 0.312 | 0.026 | 0.939 | 0.853 |
| 北京子人群 | 1973（0.435） | 0.312 | 0.048 | 0.939 | 0.853 |

[a]括号内为样本中病例与对照的样本量之比；[b]此处的遗传度为倾向阈值模型转换后的遗传度

结果显示，总人群、南京子人群和北京子人群的 SNP 遗传度估计值（标准误）分别为 0.271（0.017）、0.312（0.026）和 0.312（0.048），两个子人群的遗传度基本相同，且遗传度的标准误与样本量有关，即样本量越大，标准误越低；基于该 GWAS 研究数据估计出的 SNP 遗传度较高，在 0.3 左右，对应的 $AUC_{max}$ 和 $AUC_{half}$ 值也较大，分别超过了 0.9 和 0.8，这也提示了常见变异位点具有较强的预测能力，在肺癌发生过程中可能起到重要作用。

### （二）风险预测结果

五倍交叉验证下结果见表 2。随着-$\log_{10}(\alpha)$逐渐增大，sGRS 和 sGRS-LMM 方法的 AUC 呈现先增大后减少的趋势，这表明过于宽松的初筛水准会纳入大量的假阳性位点，而过于严格的初筛水准则会损失潜在的关联位点而降低预测的准确度；其他参数或条件相同时，各模型下 $r^2$<1 和 $r^2$<0.5 下模型预测准确度好于 $r^2$<0.2，其原因在于两种方法对 LD 结构的影响比较稳定，而高 LD 下可以提供更多的候选位点，增加了获得潜在风险位点的机会。两种方法均在 $\alpha$=1E-05 和 $r^2$<1(0.5)参数组合处获得最大的 AUC，考虑到一般情况下，$r^2$<0.5 时预测集合包含的位点数少于或等于 $r^2$<1 是预测集合内位点数，因此选择 $\alpha$=1E-05 和 $r^2$<0.5

参数下的三个模型（仅包含传统预测因素、传统因素+sGRS 位点、传统因素+sGRS-LMM 位点）进入下一步的外部数据验证分析。

表 2　交叉验证下各模型的预测准确度比较

| $-\log_{10}(\alpha)$ | LD_$r^2$ | AUC(均数±标准差) | | |
|---|---|---|---|---|
| | | 传统因素+sGRS | 传统因素+sGRS-LMM | 传统因素 [a] |
| 2 | <1 | 0.732±0.023 | 0.727±0.024 | 0.659±0.022 |
| | <0.5 | 0.729±0.007 | 0.725±0.008 | |
| | <0.2 | 0.724±0.012 | 0.722±0.011 | |
| 3 | <1 | 0.775±0.012 | 0.783±0.014 | |
| | <0.5 | 0.781±0.018 | 0.788±0.019 | |
| | <0.2 | 0.744±0.015 | 0.754±0.017 | |
| 4 | <1 | 0.805±0.011 | 0.815±0.011 | |
| | <0.5 | 0.808±0.012 | 0.819±0.011 | |
| | <0.2 | 0.775±0.011 | 0.785±0.010 | |
| 5 | <1 | 0.813±0.015 | 0.828±0.015 | |
| | <0.5 | ***0.813±0.014*** | ***0.828±0.014*** | |
| | <0.2 | 0.786±0.016 | 0.799±0.016 | |
| 6 | <1 | 0.806±0.013 | 0.825±0.012 | |
| | <0.5 | 0.806±0.013 | 0.825±0.012 | |
| | <0.2 | 0.779±0.013 | 0.794±0.014 | |
| 7 | <1 | 0.794±0.015 | 0.818±0.011 | |
| | <0.5 | 0.794±0.015 | 0.818±0.011 | |
| | <0.2 | 0.763±0.017 | 0.783±0.015 | |

[a] 传统预测因素包括：性别、年龄和吸烟（包/年）

研究策略步骤 4 的外部数据验证结果见表 3。仅包含传统因素模型（模型 1）的 AUC 为 0.687（95%可信区间：0.656~0.719）；包含传统因素和 sGRS 预测位点集合模型（模型 2）的 AUC 为 0.831（95%可信区间：0.807~0.854），sGRS 集合内 SNP 位点数为 57 个，两个模型 AUC 假设检验为 $H_0$：$AUC_1=AUC_2$，$H_1$：$AUC_1<AUC_2$，检验之 *P* 值为 1.5E-17（Bonferroni 多重校正后，下同）；包含传统因素和 sGRS-LMM 预测位点集合模型（模型 3）的 AUC 为 0.847（95%可信区间：0.824~0.869），集合内 SNP 位点数为 570,373 个，相较于模型 2，所用位点数更加庞大，但 AUC 仅提高了 0.016，单侧检验之 *P* 值为 9.7E-0.8，与模型 1 比较的 *P* 值为 1.7E-21。

表 3　预测模型的外部数据验证结果

| 预测模型 | AUC（95%可信区间） | SNP 位点数/变量数 [a] | AUC 假设检验 *P* 值 [b] |
|---|---|---|---|
| 传统因素 | 0.687（0.656~0.719） | 0/3 | |
| 传统因素+sGRS | 0.831（0.807~0.854） | 57/3 | 1.5E-17 |
| 传统因素+sGRS-LMM | 0.847（0.824~0.869） | 570,373/3 | 1.7E-21　9.7E-08 |

[a]SNP 位点数表示 sGRS 和 sGRS-LMM 两种方法构建的预测集合内位点数，变量数为用于预测的传统因素变量个数；

[b]*P* 值为三个预测模型 AUC 两两比较的假设检验结果（单侧检验，Bonferroni 多重校正），使用 DeLong 等人[27]提出的非参数方法

根据表 3 内容作三个预测模型的 ROC 曲线图，见图 2。图中，横轴为“1-特异度”，纵轴为“灵敏度”，各模型的 AUC 值即为各条曲线围成区域的面积，曲线越靠左上方，表明 AUC 越大，预测准确度越高。此外，根据模型 3 的预测得分变量作图 3，图形横轴为预测风险得分，图中虚线和实线分别为对照和病例风险得分拟合的核密度分布，近似单峰对称。两条曲线重叠越少，意味着设定的肺癌诊断截断值（临界值）可以减少正常人与病人的误判，即同时降低假阳性率和假阴性率，这与 ROC 曲线内容一致。

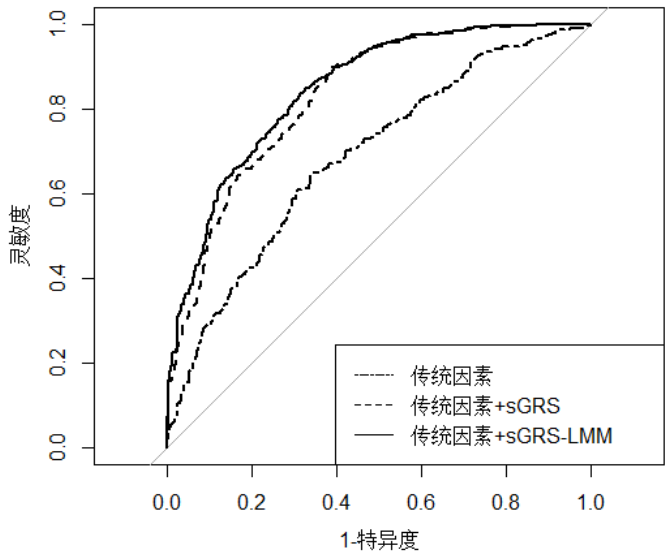


图 2　三个预测模型的 ROC 曲线图

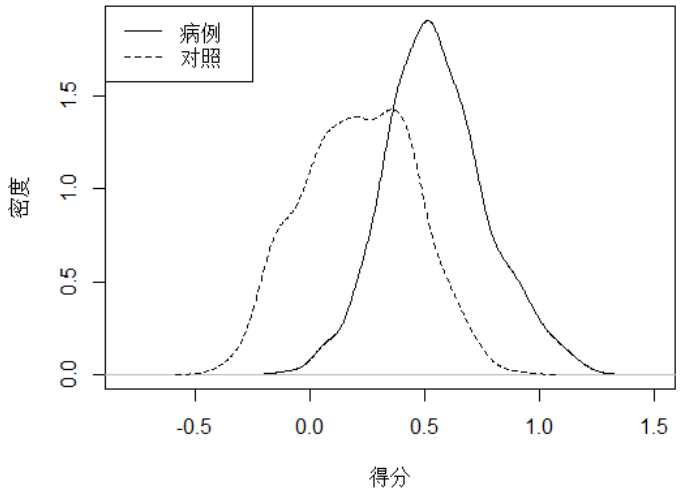


图 3　模型 3（传统因素+sGRS-LMM）病例和对照预测风险得分分布

为了考察风险得分的流行病学意义，此处进行风险得分与患病状态的关联强度分析，对风险得分进行等级化，此处使用风险得分的五分位数，评价关联强度指标则选用比值比（odds ratio，OR），结果见表 4。表格中，分别对三个模型进行分析，均以各自低风险组（即第 1 五分位）为参照，计算其他风险组相对低风险组的 OR 值及其可信区间，并进行趋势性$\chi^2$检验，结果均具有统计学意义（*P*<0.001），表示 OR 值随着风险得分的增大（风险等级增高）而增加，呈现明显的剂量-反应关系。以模型 1 为例，OR=2.24 表示高风险组（第 2 五分位）的疾病危险性为低风险组的 2.24 倍，其他含义以此类推。

表 4 各预测模型下风险得分的流行病学意义

| 模型 | 风险得分五分位数 | 病例 | 对照 | OR（95%可信区间） | 趋势性 *P* 值 [b] |
|---|---|---|---|---|---|
| 模型 1 | 0%~20% | 43 | 174 | 1.0（ref）[a] | |
| | 20%~40% | 77 | 139 | 2.24（1.42-3.55） | |
| | 40%~60% | 88 | 127 | 2.80（1.79-4.13） | *P*<0.001 |
| | 60%~80% | 110 | 106 | 4.20（2.68-6.60） | |
| | 80%~100% | 147 | 69 | 8.62（5.43-13.73） | |
| 模型 2 | 0%~20% | 10 | 206 | 1.0（ref） | |
| | 20%~40% | 42 | 174 | 4.97（2.36-11.41） | |
| | 40%~60% | 105 | 111 | 19.49（9.62-43.25） | *P*<0.001 |
| | 60%~80% | 138 | 78 | 36.45（17.85-81.01） | |
| | 80%~100% | 171 | 45 | 78.28（37.16-176.89） | |
| 模型 3 | 0%~20% | 9 | 207 | 1.0（ref） | |
| | 20%~40% | 46 | 170 | 6.22（2.89-14.83） | |
| | 40%~60% | 95 | 121 | 18.06（8.65-41.93） | *P*<0.001 |
| | 60%~80% | 137 | 79 | 39.89（18.99-92.60） | |
| | 80%~100% | 179 | 37 | 111.27（50.50-264.75） | |

[a]ref 为参考组；[b]趋势性$\chi^2$检验之 *P* 值

组成模型 2 遗传部分的 sGRS 预测集合共包含 57 个位点，其中 *rs*1846594 和 *rs*2736100 两个位点在原肺癌关联研究中已经报道过，还有 11 个位点所在的基因可能与癌症（肺癌）的发生、代谢相关，剩余位点均是位于目前尚未有证据证明与癌症有关的基因上或是基因间区内。这 11 个位点的位置、所在基因以及基因的生物学功能见表 5。

表 5　sGRS 方法下预测集合的位点信息

| SNP 编号 | 染色体 | 位置 | 基因 | 基因功能 |
|---|---|---|---|---|
| *rs*2469099 | 15 | 64702619 | 基因间区，毗邻 KIAA0101 | 高表达与食管癌、胃癌和肝癌等有关[35, 36] |
| *rs*1376910 | 5 | 102439563 | SLCO6A1 | 编码有机阴离子转运蛋白家族，在非小细胞肺癌细胞中存在高表达[37] |
| *rs*17115702 | 14 | 72442397 | RGS | 胰腺癌、乳腺癌相关[38, 39] |
| *rs*2400795 | 5 | 101803238 | SLCO6A1 | 抑制甲状腺肿瘤（促进衰老、抑制侵袭）、肺癌[40-42] |
| *rs*1872645 | 12 | 24814278 | BCAT1 | 鼻咽癌、结直肠癌[43, 44] |
| *rs*12544802 | 8 | 142012928 | PTK2（FAK） | 影响癌症发生[45] |
| *rs*1178120 | 7 | 18714500 | HDAC9 | 肺癌发生相关[46] |
| *rs*3743072 | 15 | 76708817 | CHRNB4 | 尼古丁依赖[47] |
| *rs*465498 | 5 | 1325688 | CLPTM1L | 肺癌发生相关[48] |
| *rs*10035133 | 5 | 102649969 | LINC00491 | 胰腺癌相关[49] |
| *rs*6712250 | 2 | 132589162 | GPR39 | 食管癌相关[50] |

# 三、讨论

## （一）复杂疾病的风险预测研究

目前的风险预测研究表明仅使用经过 GWAS 严格验证的 SNP 位点进行预测，其效果难以令人满意：Mealiffe 等人[51]应用大型临床试验的乳腺癌病例和对照进行实例分析，发现 7 个 SNPs 位点预测的 AUC 为 0.587；Ripatti 等[52]的一项前瞻性研究利用早期七个研究发现的与冠心病、心肌梗死等疾病相关的 13 个 SNPs 位点构建遗传风险得分，相较于传统风险因素和家族史，并不能提高 AUC 值；Johansson 等人[53]的研究发现 33 个 SNPs 位点仅能在前列腺特异抗原的基础上把 AUC 值提高了 0.01；还有很多类似的研究[54, 55]也在提示相同的信息。GWAS 研究的多阶段验证虽然保证了一定的真实性和可靠性，但其代价在于忽略了可能大量存在的低效应位点。针对这一点，后来的一些改进研究主要演变为两类策略：

一类是放宽 $\alpha$ 水准，与此同时考虑 LD 结构的影响进行位点修剪，如 Wei 等人[6]的 I 型糖尿病研究；另一类是基于全基因组 SNP 的混合效应模型，即 BLUP 及其相关改进方法。而本研究也正是基于这两类策略继续开展风险预测研究。

## （二）肺癌的风险预测研究与方法评价

本研究中，仅利用传统因素的预测准确度（AUC）为 0.687，结合 sGRS 方法寻找的 SNP 位点集（共 57 个）将 AUC 提高到 0.831，而结合 sGRS-LMM 方法的全部位点将 AUC 进一步提高到 0.847，但这些值都低于估计的理论 AUC 值（>0.9），与 $AUC_{half}$ 近似，其可能原因在于：位点间关系复杂，sGRS 方法所找到的位点集很有可能并非最佳的预测子集，而 $\alpha$=1E-05 和 $r^2$<0.5 也仅是预设参数下的最佳组合并非全局最佳（设定过多的参数组合情况会极大的增加计算时间），这需要通过方法改进来解决。此外，训练集样本量较小、真实关联位点与 SNP 之间不完全的 LD 结构也可能是比较重要的因素。对比同类研究，Li 等人[56]利用 GWAS 严格验证的 4 个肺癌关联位点和吸烟史对中国人群肺癌风险预测，AUC 仅为 0.639，而单纯使用吸烟因素则为 0.619。

在 sGRS 预测集合中，许多位点所在基因的功能与癌症的发生发展有一定的关联，如 *rs*1376910 所在的基因 SLCO6A1，在非小细胞肺癌细胞中存在高表达，而 *rs*3743072 所在的基因 CHRNB4 则与尼古丁依赖有关，进而可能通过吸烟因素来影响肺癌发生。当然，我们也要认识到这些位点可能仅仅是易感位点的标记物，而寻找这些真正的位点则需要更多的研究。

本研究针对 SNP 位点使用了 sGRS 和 sGRS-LMM 两种方法。sGRS 的优势在于可以获得较高的预测准确度，对应的位点数目一般较少，临床实践应用花费少有利于推广，从其方法构造的角度，它对 LD 结构更加稳定；sGRS-LMM 则将 sGRS 未纳入位点的效应作为随机效应，充分利用所有位点，因此预测效果会好于前者，但因为剩余的潜在风险位点效应较低、数量较少，其提高的绝对差值通常有限（疾病间遗传模式差异大，结果可能不尽相同），这也就存在分型了更庞大数量的位点（意味着更大的经济预算）与预测准确度提高有限的矛盾。因此方法的选择通常需要根据研究、应用的目的和条件来决定。基于 GWAS 数据的预测研究的主要缺点在于计算时间较长和占用巨大的内存，这同样也是本研究两种方法的限制之一。从研究的过程来看，交叉验证筛选 $\alpha$ 和 $r^2$ 过程耗费了较长的时间，但就相关研究[8]的和本研究的结果，通常而言初筛界值 $\alpha$=1E-05（$r^2$ 不应过低）预测效果较好（虽然可能并非最佳）可以推荐直接使用，而不需要再通过交叉验证步骤来筛选，如果想要追求更高的预测准确度，则可以先进行交叉验证筛选。

## （三）遗传度估计

本研究根据 GWAS 数据估计了中国非小细胞肺癌的 SNP 遗传度，主要目的在于指导方法的预测准确度评价。遗传度估计值为 0.3 左右，标准误随样本量的增大而变小。如果后续的研究可以考虑其他的加性遗传因素（如罕见变异的位点），那么狭义遗传度的值很有可能大于 0.3。国内有研究者[57]通过核心家系估计肺癌的狭义遗传度为 0.246，其中女性人群达到 0.378；另一研究[58]也通过家系数据估计苏州地区吸烟者和非吸烟者遗传度分别为 0.406±0.040、0.276±0.048；国外有研究者[24, 59]通过双生子研究流行病学参数估计肺癌遗传度约为 0.76。由此可见，不同的人群其遗传度亦不相同，而较低的遗传度往往限制了模型中遗传部分的预测效果。因此，在评价遗传研究中预测模型的优劣时，应首先认清楚对应性状（疾病）的遗传度大小。以本研究为例，利用所有加性效应遗传因素进行预测研究的理论 AUC 可以达到 0.9 以上。需要说明的是，虽然本研究的遗传方差是指可以遗传的加性方差，而且是常见变异位点所解释的加性方差，但就遗传风险预测研究本身而言，并不应该拘泥于此，考虑位点间或基因间交互作用的上位性方差亦会贡献预测准确度。

# 四、总结

本研究的特点在于提供了一个基于 GWAS 数据和传统因素的多步骤风险预测分析流程，并提出了适合高维数据的风险预测方法：相对于传统因素，SNP 位点可以较大幅度地提高预测的 AUC 值，表明该 GWAS 中的常见变异位点是肺癌的强预测因子；sGRS 和 sGRS-LMM 方法可以用于肺癌全基因组关联研究数据的疾病风险预测，前者所需的预测位点数较少，有利于公共卫生或临床实践应用，而后者预测准确度更高，适合基于全部常见变异位点的预测研究；与此同时，考虑到交叉验证中过程筛选参数的计算耗时，本研究给出了 $\alpha$ 的推荐值。

本研究的 AUC 值相较理论值尚有差距，这也意味着预测模型还有较大的提高空间，如可以从增大样本量、改进预测方法等角度实施。但就基于 GWAS 的预测研究而言，并不应该拘泥于常见变异位点效应的简单累加，考虑位点间或基因间交互作用、基因-环境交互作用亦会贡献预测准确度。因此，针对 GWAS 数据的风险预测研究还将继续延续下去，这也可以为以后基于测序数据（提供罕见变异和更多的常见变异位点）、多组学（Multi-Omics）数据的风险预测提供指导。

# 参考文献

1. Mahmood S S, Levy D, Vasan R S, et al. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective[Z]. 2014: 383, 999-1008.
2. Catalog of Published Genome-Wide Association Studies[Z]. 2015: 2015.
3. van der Net J B, Janssens A C, Sijbrands E J, et al. Value of genetic profiling for the prediction of coronary heart disease. Am Heart J. 2009, 158(1): 105-110.
4. Mihaescu R, Meigs J, Sijbrands E, et al. Genetic risk profiling for prediction of type 2 diabetes. PLoS Curr. 2011, 3: N1208.
5. Cook N R. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007, 115(7): 928-935.
6. Wei Z, Wang K, Qu H Q, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. 2009, 5(10): e1000678.
7. de Maturana E L, Chanok S J, Picornell A C, et al. Whole genome prediction of bladder cancer risk with the Bayesian LASSO. Genet Epidemiol. 2014, 38(5): 467-476.
8. Speed D, Balding D J. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014, 24(9): 1550-1557.
9. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 2013, 9(2): e1003264.
10. Hu Z, Wu C, Shi Y, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese[Z]. 2011: 43, 792-796.
11. Visscher P M, Hill W G, Wray N R. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet. 2008, 9(4): 255-266.
12. Yang J, Benyamin B, Mcevoy B P, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010, 42(7): 565-569.
13. Yang J, Lee S H, Goddard M E, et al. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011, 88(1): 76-82.
14. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008, 456(7218): 18-21.
15. Gilmour A R, Thompson R, Cullis B R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics. 1995: 1440-1450.
16. Johnson D L, Thompson R. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. Journal of dairy science. 1995, 78(2): 449-456.
17. Harville D A. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association. 1977, 72(358): 320-338.
18. Lynch M, Walsh B. Genetics and analysis of quantitative traits[M]. Sinauer Sunderland, 1998.

19. Lange K, Westlake J, Spence M. Extensions to pedigree analysis III. Variance components by the scoring method. Annals of human genetics. 1976, 39(4): 485-491.
20. Falconer D S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. Annals of Human Genetics. 1965, 29(1): 51-76.
21. Dempster E R, Lerner I M. Heritability of threshold characters. Genetics. 1950, 35(2): 212.
22. Van Vleck L D. Estimation of heritability of threshold characters. Journal of dairy science. 1972, 55(2): 218-225.
23. Lee S H, Wray N R, Goddard M E, et al. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011, 88(3): 294-305.
24. Wray N R, Yang J, Goddard M E, et al. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. 2010, 6(2): e1000864.
25. Mason S J, Graham N E. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society. 2002, 128(584): 2145-2166.
26. Mann H B, Whitney D R. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics. 1947: 50-60.
27. Delong E R, Delong D M, Clarke-Pearson D L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988, 44(3): 837-845.
28. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011, 12: 77.
29. Kang H M, Sul J H, Service S K, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010, 42(4): 348-354.
30. Robinson G K. That BLUP is a good thing: the estimation of random effects. Statistical science. 1991: 15-32.
31. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007, 81(3): 559-575.
32. 陈万青，张思维，郑荣寿，等. 中国2009年恶性肿瘤发病和死亡分析. 中国肿瘤. 2013(01): 2-12.
33. 韩仁强，郑荣寿，张思维，等. 1989年-2008年中国肺癌发病性别、城乡差异及平均年龄趋势分析. 中国肺癌杂志. 2013(09): 445-451.
34. 陈万青，张思维，曾红梅，等. 中国2010年恶性肿瘤发病与死亡. 中国肿瘤. 2014(01): 1-10.
35. Su X, Zhang T, Cheng P, et al. KIAA0101 mRNA overexpression in peripheral blood mononuclear cells acts as predictive marker for hepatic cancer. Tumor Biology. 2014, 35(3): 2681-2686.
36. Cheng Y, Li K, Diao D, et al. Expression of KIAA0101 protein is associated with poor survival of esophageal cancer patients and resistance to cisplatin treatment in

vitro. Laboratory Investigation. 2013, 93(12): 1276-1287.
37. Brenner S, Klameth L, Riha J, et al. Specific expression of OATPs in primary small cell lung cancer (SCLC) cells as novel biomarkers for diagnosis and therapy. Cancer letters. 2015, 356(2): 517-524.
38. Jiang N, Xue R, Bu F, et al. Decreased RGS6 expression is associated with poor prognosis in pancreatic cancer patients. Int J Clin Exp Pathol. 2014, 7(7): 4120-4127.
39. Maity B, Yang J, Huang J, et al. Regulator of G protein signaling 6 (RGS6) induces apoptosis via a mitochondrial-dependent pathway not involving its GTPase-activating protein activity. J Biol Chem. 2011, 286(2): 1409-1419.
40. Terauchi K, Shimada J, Uekawa N, et al. Cancer-associated loss of TARSH gene expression in human primary lung cancer. J Cancer Res Clin Oncol. 2006, 132(1): 28-34.
41. Latini F R, Hemerly J P, Oler G, et al. Re-expression of ABI3-binding protein suppresses thyroid tumor growth by promoting senescence and inhibiting invasion. Endocrine-related cancer. 2008, 15(3): 787-799.
42. Uekawa N, Terauchi K, Nishikimi A, et al. Expression of TARSH gene in MEFs senescence and its potential implication in human lung cancer. Biochem Biophys Res Commun. 2005, 329(3): 1031-1038.
43. Zhou W, Feng X, Ren C, et al. Over-expression of BCAT1, a c-Myc target gene, induces cell proliferation, migration and invasion in nasopharyngeal carcinoma. Mol Cancer. 2013, 12: 53.
44. Yoshikawa R, Yanagi H, Shen C S, et al. ECA39 is a novel distant metastasis-related biomarker in colorectal cancer. World J Gastroenterol. 2006, 12(36): 5884-5889.
45. Gabarra-Niecko V, Schaller M D, Dunty J M. FAK regulates biological processes important for the pathogenesis of cancer. Cancer and Metastasis Reviews. 2003, 22(4): 359-374.
46. Okudela K, Mitsui H, Suzuki T, et al. Expression of HDAC9 in lung cancer--potential role in lung carcinogenesis. Int J Clin Exp Pathol. 2014, 7(1): 213-220.
47. Haller G, Druley T, Vallania F L, et al. Rare missense variants in CHRNB4 are associated with reduced risk of nicotine dependence. Human molecular genetics. 2012, 21(3): 647-655.
48. Li C, Yin Z, Wu W, et al. Genetic variants in TERT-CLPTM1L genetic region associated with several types of cancer: a meta-analysis. Gene. 2013, 526(2): 390-399.
49. Li J, Liu D, Hua R, et al. Long non-coding RNAs expressed in pancreatic ductal adenocarcinoma and lncRNA BC008363 an independent prognostic factor in PDAC. Pancreatology. 2014, 14(5): 385-390.
50. Xie F, Liu H, Zhu Y H, et al. Overexpression of GPR39 contributes to malignant development of human esophageal squamous cell carcinoma. BMC Cancer. 2011, 11: 86.
51. Mealiffe M E, Stokowski R P, Rhees B K, et al. Assessment of clinical validity of

a breast cancer risk model combining genetic and clinical information. J Natl Cancer Inst. 2010, 102(21): 1618-1627.
52. Ripatti S, Tikkanen E, Orho-Melander M, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet. 2010, 376(9750): 1393-1400.
53. Johansson M, Holmstrom B, Hinchliffe S R, et al. Combining 33 genetic variants with prostate-specific antigen for prediction of prostate cancer: longitudinal study. Int J Cancer. 2012, 130(1): 129-137.
54. Herder C, Roden M. Genetics of type 2 diabetes: pathophysiologic and clinical relevance. Eur J Clin Invest. 2011, 41(6): 679-692.
55. de Miguel-Yanes J M, Shrader P, Pencina M J, et al. Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. Diabetes Care. 2011, 34(1): 121-125.
56. Li H, Yang L, Zhao X, et al. Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model. BMC Med Genet. 2012, 13: 118.
57. 金永堂，周晓铁，何兴舟. 宣威肺癌遗传因素作用大小的估算与分析. 中国肺癌杂志. 2001, 4(5): 354-356.
58. 郭志荣，彭盛梅，蒋国雄，等. 肺癌的遗传流行病学研究. 中华预防医学杂志. 1996(03): 28-30.
59. Risch N. The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. Cancer Epidemiol Biomarkers Prev. 2001, 10(7): 733-741.