arXiv:2601.02583v1 [stat.ME] 5 Jan 2026

利用协变量注释信息提高 Knockoff 方法的检验效能

Xiangyu Zhang[1],Lijun Wang[2],Changjun Li[1],Chen Lin[1],Hongyu Zhao[1*]

1 生物统计系，公共卫生学院，耶鲁大学，纽黑文，康涅狄格州，美利坚合众国**2** 统计系，数学科学学院，浙江大学，杭州，浙江，中国

*通讯作者：hongyu.zhao@yale.edu

摘要

全基因组关联研究（GWAS）常常在基因组区域内发现众多遗传变异与感兴趣性状之间的关联信号。这些变异的功能注释提供了有价值的先验信息，有助于优先识别具有生物学意义的遗传变异，并提高检测因果变异的检验效能。然而，由于这些变异之间存在显著相关性，如何在有效利用先验知识的同时严格控制假发现率，仍是一个关键问题。我们提出了注释知情敲除（AnnoKn），这是一种基于 Knockoff 的方法，能够在严格控制假发现率的前提下进行注释信息指导的变量选择。AnnoKn 将 knockoff程序与自适应 Lasso 回归相结合，在统一的贝叶斯方法框架下纳入功能注释信息，以评估多个协变量的重要性。为促进个体水平数据不可获取时的实际应用，我们进一步将 AnnoKn 扩展至可使用汇总统计量运行。通过模拟以及在 GTEx 和 GWAS 数据集上的真实应用，我们表明，与现有的注释信息指导的变量选择方法相比，AnnoKn 在保持对假发现有效控制的同时，能够以更高的检验效能检测因果遗传变异。

1 引言

全基因组关联研究（GWASs）在检测与复杂人类性状相关的遗传变异方面取得了显著成功 [1]。其他遗传研究——称为全转录组关联研究（TWASs） [2, 3], ——也揭示了人类疾病背后的分子调控机制。然而，传统关联研究只评估单个协变量–响应变量对之间的边际关联，而忽略了基因组区域内遗传因素之间的大量相关性 [4, 5]。在高维

基因组分析中，协变量往往呈现复杂的相关结构，例如连锁不平衡（LD），这使得区分真正的因果信号与相关的非因果协变量变得困难 [6]。为克服这一挑战并识别真正的因果信号，我们可以纳入关于协变量的外部信息，或明确建模其潜在的依赖结构 [7, 8, 9]。

随着生物医学数据的数量和多样性迅速增长，研究人员对有助于基因组分析和科学发现的遗传变异有了更多了解。一个例子是表达数量性状位点（eQTLs）的鉴定，它们是影响基因表达水平的遗传变异 [10, 11]。研究表明，功能注释——例如调控元件、染色质可及性或转录因子结合位点——可以优先筛选具有真实调控效应的遗传变异，从而提升表达数量性状位点检测的检验效能和准确性 [12, 13, 14]。另一个例子是全基因组关联研究数据的分析：在目标人群中开展关联研究之前，可以利用其他人群的汇总统计量（p-值或 Z-分数）作为注释信息，优先筛选风险遗传变异 [15]。因此，亟需开发能够利用功能注释提升检验效能，同时严格控制相关协变量中的假发现率（FDR）的统计程序，从而确保注释信息辅助发现的可靠性。

已有多种推断方法被提出，用于识别有信息量的协变量，同时保证错误发现率控制。在所有假设都具有有效 p值的经典设定下，当检验统计量相互独立或满足某些形式的正依赖时，Benjamini–Hochberg（BH）程序可控制错误发现率 [16, 17]。近来，研究者开发了自适应 p值阈值法（AdaPT），利用可用的协变量先验信息，构造一系列逐渐严格的拒绝阈值 [18]。然而，标准的基于 p值的错误发现率控制方法需要已知有效的 p值，并且要求其依赖关系满足特定假设 [19], ，由此带来两个问题。首先， p值仅评估边际关联，因此所得发现难以解释：许多与真实信号相关但并不重要的协变量仍可能表现出很强的边际相关性 [20]。其次，尽管已有多种 BH 程序的改进方法被提出，以便在一般依赖结构下控制错误发现率，但这些方法往往较为保守，通常会降低统计功效 [21]。这一问题在遗传学中尤为突出，因为协变量（如单核苷酸多态性（SNP））之间普遍存在复杂的相关结构 [22]。

敲除框架避免直接使用 p 值，并提供了一种可靠的方法来进行受控变量选择，而无需对响应变量的条件分布作出假设 [23, 24]。通过对协变量的相关结构建模，model-X Knockoff 旨在最大限度地发现条件依赖协变量，同时严格控制名义水平下的假发现率。基于敲除框架，自适应 knockoff 筛选器（AdaKn）被提出，用于纳入先验注释信息 [25]。AdaKn 首先应用敲除框架，在数据驱动的拒绝路径上构建假设的初始排序。然后，它利用先验信息来优化这一排序

将先验支持较强的协变量移至序列前端，从而提高更早发现真实信号的可能性，同时维持错误发现率控制。然而，AdaKn 本质上是一种两阶段方法，其中第一阶段构建协变量重要性度量时未能有效纳入先验信息。这可能会限制整体统计功效。此外，AdaKn 是为响应变量和协变量均被完整观测的情形而设计的，这限制了其实际适用性，因为在许多现实场景中，能够获得的只是每对响应变量与协变量之间边际关联的汇总统计量，而非个体水平数据 [26, 27]。

我们提出了注释知情 Knockoff（AnnoKn），这是一种基于 Knockoff 的框架，可自适应地利用功能注释，在严格控制假发现率（FDR）的同时提升 Knockoff 方法的检验效能。与两阶段方法不同，AnnoKn 在统一的贝叶斯框架中整合了 Knockoff 程序和先验信息。AnnoKn 利用功能注释等协变量的先验知识，自适应地学习特定于协变量的惩罚项，以最大化后验似然。此外，我们进一步扩展了 AnnoKn，使其能够在无法获取个体水平数据的情况下，结合汇总统计量和估计得到的协变量两两相关矩阵进行分析。通过广泛的模拟和真实数据分析（包括 eQTL 定位，以及身高和精神分裂症的全基因组关联研究），我们证明，与现有方法相比，AnnoKn 在检测因果遗传变异时检验效能更高，同时仍能有效控制假发现率。此外，AnnoKn 能有效处理高维注释，并自动滤除非信息性注释。

2 方法

2.1 model-X Knockoff 和 AdaKn 概述

y 表示响应变量，$\boldsymbol{X} = (X_1, \ldots\ldots, X_p)$ 表示 p 个协变量。为了推断哪些协变量确实与响应变量相关，model-X Knockoffs [24] 对零假设进行条件检验

$$H_0 : X_j \perp y | X_{-j}, \quad j = 1, \ldots, p,$$

其中 X_{-j} 表示除第 j 个协变量外的所有协变量集合。

要实现 model-X Knockoffs，首先必须构造 knockoff 副本，记为 $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots\ldots, \tilde{X}_p)$。这些随机变量的设计旨在满足以下两个条件：

1. $(\boldsymbol{X}, \tilde{\boldsymbol{X}})_{\text{swap}(S)} \overset{\text{d}}{=} (\boldsymbol{X}, \tilde{\boldsymbol{X}}),\ \forall \text{ subset } S \subset \{1, \ldots, p\}.$
2. $\tilde{\boldsymbol{X}} \perp y | \boldsymbol{X}.$

借助 knockoff 副本，为每个协变量计算一个特征重要性统计量 W_j。这些特征统计量旨在满足两个关键性质：（1）交换第 j 个协变量及其 knockoff 副本，会改变 W_j 的符号（翻转符号性质）；（2）

W_j 对于非零假设协变量，往往取较大的正值。给定一个固定阈值 $t > 0$，选择 $W_j \geq t$ 的协变量作为显著协变量。根据翻转符号性质，与零假设协变量对应的统计量 W_j 关于零对称，这意味着

$$\#\{j : W_j \leq -t\} \geq \#\{\text{null } j : W_j \leq -t\} \stackrel{\text{d}}{=} \#\{\text{null } j : W_j \geq t\}.$$

根据这个不等式，可以估计假发现比例（FDP）：

$$\text{FDP}(t) = \frac{\#\{\text{null } j : W_j \geq t\}}{\#\{j : W_j \geq t\}} \leq \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}},$$

并且可以将显著特征统计量的阈值 T 选为

$$T = \min\left\{t : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\},$$

其中， q 是目标 FDR 水平。对于每个协变量 X_i，与其特征统计量 W_i 相关联的 q-值 q_i 定义如下:

$$q_i = \min_{t \leq T_i} \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}.$$

通过找出相应的 q-值 q_i 满足 $q_i < q$ 的协变量，即可在目标水平 q 下控制 FDR 并完成协变量筛选。

model-X Knockoff 的变量选择过程如下。首先，按照特征统计量的绝对值从小到大对协变量排序:

$$|W_{\pi_1}| \leq \cdots \leq |W_{\pi_k}| \leq \ldots |W_{\pi_p}|.$$

对于从 $k = 0$ 开始的每一步 $k \in \{0, 1, \ldots, p\}$，该方法会考虑选择所有满足 $j > k$ 和 $W_{\pi_j} > 0$ 的假设 π_j ，并计算相应的假发现比例估计值。当估计的假发现比例低于目标 FDR 水平时，过程在最小的 k 处停止。

给定这一假设有序序列，AdaKn [25]会调整其顺序，将先验支持更强的协变量移至序列末尾，从而提高它们被选中的可能性。在每一步 k，AdaKn 都会使用自适应 knockoff 筛选器 ϕ_{k+1} 来确定序列中的下一个假设 π_{k+1} 。只要 ϕ_{k+1} 完全由 $\{|W_j|\}_{j\in\{1,\ldots,p\}}$， $\{W_{\pi_j}\}_{j>k}$ 和可用的先验信息决定，就可以证明错误发现率控制成立。

AdaKn 使用多种自适应筛选器实现这一自适应 knockoff 框架，包括广义线性模型（GLM）、广义加性模型（GAM）、随机森林（RF）和贝叶斯双组模型（期望最大化）筛选器，这些筛选器均纳入了我们的比较。

2.2 F基于 Knockoffs 和自适应 Lasso 特征统计量的错误发现率控制

我们不采用上述用于纳入先验信息的两步流程，而是在统一框架内使用自适应 Lasso 回归 [28]计算量化协变量重要性的特征统计量，并利用注释信息对其进行优先排序：

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{2p}} l(y,[\boldsymbol{X},\tilde{\boldsymbol{X}}];\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}\in\mathbb{R}^{2p}} \frac{1}{2}\|y-[\boldsymbol{X},\tilde{\boldsymbol{X}}]\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{p}\phi_j(|\beta_j|+|\beta_{j+p}|), \tag{1}$$

其中 $\boldsymbol{\beta}=(\beta_1, \ldots\ldots, \beta_p, \beta_{p+1}, \ldots\ldots, \beta_{2p})^T$是系数向量（前 p 个分量对应原始协变量，后 p 个分量对应 Knockoffs），λ 是正则化超参数，ϕ_j 是第 j 个协变量的先验重要性权重。

用以并将系数 $\hat{\boldsymbol{\beta}}$ 的绝对值定义问题的解：$Z_j=|\beta_j|$和 $Z_j=|\beta_{j+p}|$，我们为第 j 个协变量构造 Lasso 系数差（LCD）特征统计量 W_j：

$$W_j([\boldsymbol{X},\tilde{\boldsymbol{X}}],y) = Z_j - \tilde{Z}_j = |\hat{\beta}_j| - |\hat{\beta}_{j+p}|.$$

如文献 [24], 所述，若特征统计量 W_j 满足翻转符号性质，则可确保有效的假发现率控制。我们在定理1中验证了所提出统计量的这一性质。

定理1（自适应Lasso LCD统计量的翻转符号性质）。设 W_j 为由式（1）中定义的自适应 Lasso 回归得到的 LCD统计量。对于任意索引子集 $S\subset\{1,...,p\}$，以下翻转符号性质成立：

$$w_j\big([\boldsymbol{X},\tilde{\boldsymbol{X}}]_{\mathrm{swap}(S)},y\big) = \begin{cases} w_j\big([\boldsymbol{X},\tilde{\boldsymbol{X}}],y\big), & j\notin S,\\ -w_j\big([\boldsymbol{X},\tilde{\boldsymbol{X}}],y\big), & j\in S,\end{cases}$$

其中，$[\boldsymbol{X},\boldsymbol{X}]_{\mathrm{swap}(S)}$是通过对每个索引$j\in S$交换 $[\boldsymbol{X},\boldsymbol{X}]$ 中的条目 X_j和 X_j得到的。

The 定理1的证明见补充材料第S3.1节.

2.3 利用个体水平数据的注释信息构建 Knockoff

假设我们获得了 n 个独立同分布观测 $(X_{i1},...,X_{ip},y_i)$。令 $\mathbf{y}=[y_1,\ldots,y_n]^\top$为个样本的观测响应，$\mathbf{X}\in\mathbb{R}^{n\times p}$ 表示设计矩阵。令$\mathbf{X}$ 为构造的敲除矩阵。在本研究中，我们假设响应向量$\mathbf{y}$、设计矩阵$\mathbf{X}$和敲除矩阵$\mathbf{X}$ 均已标准化，使所有协变量（列）的均值为零、方差为单位方差。为定义自适应 Lasso 的注释信息引导的惩罚权重ϕ_j，我们将线性回归模型嵌入以下贝叶斯层级框架：

$$\mathbf{y} = [\mathbf{X},\tilde{\mathbf{X}}]\boldsymbol{\beta}+\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon}\sim\mathcal{N}(0,\sigma^2\mathbf{I}_n),$$

$$\beta_j = \beta_{j+p} \sim Laplace(0, (n\lambda_0\phi_j)^{-1}), \quad j = 1, ..., p,$$

$$\phi_j = \exp\left(\sum_{l=1}^{L} \lambda_l A_{jl}/d\right), \quad \lambda_l \sim \mathcal{N}(0, \tau^2), \tag{2}$$

其中$\mathbf{A} = \{A_{jl}\}$ 为标准化后的 $p \times L$ 注释矩阵， λ_0 为预先选定的总体正则化参数， σ^2 为误差方差超参数， d控制注释权重 λ_l的缩放或收敛速度， τ^2 为注释权重 λ_l的先验方差超参数。

将 $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_L)$ 记为目标参数向量，将 $\boldsymbol{\theta} = (\lambda_0, \sigma^2, d, \tau^2)$ 记为超参数向量，则后验概率可计算如下:

$$\begin{aligned}
&p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{A}, \mathbf{y}, \boldsymbol{\theta}) \\
\propto &p(\mathbf{y}|\mathbf{X}, \tilde{\mathbf{X}}, \boldsymbol{\beta}, \boldsymbol{\theta}) \times p(\boldsymbol{\beta}|\mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\theta}) \times p(\boldsymbol{\lambda}|\boldsymbol{\theta}) \\
= &(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta}\|^2\right) \\
&\times \prod_{j=1}^{p} \left(\frac{n\lambda_0\phi_j}{2}\right) \exp(-n\lambda_0\phi_j|\beta_j|) \times \prod_{j=1}^{p} \left(\frac{n\lambda_0\phi_j}{2}\right) \exp(-n\lambda_0\phi_j|\beta_{j+p}|) \\
&\times (2\pi\tau^2)^{-L/2} \exp\left(-\frac{1}{2\tau^2}\sum_{l=1}^{L}\lambda_l^2\right) \\
\propto &\exp(-\frac{1}{2\sigma^2}\|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta}\|_2^2) \times \prod_{j=1}^{p} \phi_j^2 \exp(-n\lambda_0\phi_j(|\beta_j| + |\beta_{j+p}|)) \times \exp\left(-\frac{1}{2\tau^2}\sum_{l=1}^{L}\lambda_l^2\right).
\end{aligned}$$

˜鉴于已将 $[\mathbf{X}, \mathbf{X}]$ 和 $\mathbf{y}$ 标准化为单位方差，我们将误差方差超参数固定为 $\sigma^2 = 1$ ，而不对其进行估计。这样简化有两个理由。首先，在多基因性状的情境下，当我们将分析范围限制在某个基因组区域内时，表型方差的大部分来自随机误差，而非遗传效应（残差方差 ≈ 1）。其次，根据目标函数的结构，σ^2 可以有效地并入正则化参数 λ_0 和 τ^2。因此，固定 σ^2 不会影响优化路径，也不会影响目标参数 $\boldsymbol{\beta}$和 $\boldsymbol{\lambda}$的选择。由此得到的对数后验为:

$$\begin{aligned}
&\log p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{A}, \mathbf{y}, \boldsymbol{\theta}) \\
= &-\frac{1}{2}\|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta}\|_2^2 + 2\sum_{j=1}^{p}\log\phi_j - n\lambda_0\sum_{j=1}^{p}\phi_j(|\beta_j| + |\beta_{j+p}|) - \frac{1}{2\tau^2}\sum_{l=1}^{L}\lambda_l^2 + \text{const.}
\end{aligned} \tag{3}$$

请注意，当固定 $\boldsymbol{\lambda}$ 时，相对于 $\boldsymbol{\beta}$ 最大化式 (3) 中的对数后验，等价于求解式 (1) 中的自适应 Lasso 模型。因此，我们提出一种交替最大化策略，在固定 λ_0的条件下，针对 $\boldsymbol{\beta}$ 和 $\boldsymbol{\lambda}$优化目标对数后验：

- 步骤 1：我们初始化 $\phi_j = 1$ 的先验权重 $j = 1, \ldots, p$，并将初始的 $\boldsymbol{\beta}$设为普通最小二乘法 (OLS) 估计值。
- 步骤 2：给定 ϕ_j的当前估计值，我们通过求解自适应 Lasso 问题来更新 $\boldsymbol{\beta}$ ：

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta}\|_2^2 + \lambda_0 \sum_{j=1}^{p} \phi_j (|\beta_j| + |\beta_{j+p}|) \right\}.$$

- 步骤 3：给定 $\boldsymbol{\beta}$的当前估计值，并利用 $\mathbf{A}$ 已标准化这一事实，即对于 $l \in \{1, \ldots, L\}$, $\sum_{j=1}^{p} A_{jl} = 0$，我们通过最大化以下关于 $\boldsymbol{\lambda}$的约简表达式来更新 $\hat{\boldsymbol{\lambda}}$：

$$\begin{aligned}
&2\sum_{j=1}^{p} \log \phi_j - n\lambda_0 \sum_{j=1}^{p} \phi_j (|\beta_j| + |\beta_{j+p}|) - \frac{1}{2\tau^2} \sum_{l=1}^{L} \lambda_l^2 \\
=&2\sum_{j=1}^{p} \sum_{l=1}^{L} \lambda_l A_{jl}/d - n\lambda_0 \sum_{j=1}^{p} \exp(\sum_{l=1}^{L} \lambda_l A_{jl}/d)(|\beta_j| + |\beta_{j+p}|) - \frac{1}{2\tau^2} \sum_{l=1}^{L} \lambda_l^2 \\
=&-n\lambda_0 \sum_{j=1}^{p} \exp(\sum_{l=1}^{L} \lambda_l A_{jl}/d)(|\beta_j| + |\beta_{j+p}|) - \frac{1}{2\tau^2} \sum_{l=1}^{L} \lambda_l^2, \qquad (4)
\end{aligned}$$

  并基于更新后的 $\boldsymbol{\lambda}$估计值更新 ϕ_j 。

- 步骤 4（迭代）：在步骤 2 和步骤 3 之间迭代，直至收敛，然后输出最终估计值 $\boldsymbol{\beta}$ 和 $\boldsymbol{\lambda}$。

基于这一策略，我们提出算法 1，在候选集 $\mathbb{A}$中调整超参数 λ_0，并选择表现最佳的 λ_0 ，以最小化交叉验证误差。

尽管算法 1 的表现良好，但由于其计算开销取决于候选集的大小 $\mathbb{A}$ for λ_0，计算成本可能很高。因此，我们提出算法 2。研究表明，算法 2 的表现与算法 1 相近，同时计算效率显著更高。

算法 1 注释知情敲除（AnnoKn）的主算法

需要： 响应向量$\mathbf{y}$、协变量矩阵$\mathbf{X}$、敲除副本$\tilde{\mathbf{X}}$、注释矩阵$\mathbf{A}$，超参数 d，$\tau^2$1: **初始化：** $\lambda_0^f = 0$, $c^f = \infty$, $\phi_j^f = 1$ 对于 $j = 1, \ldots,$ p2: **对** λ_0 中的每个 $\mathbb{A}$ **执行**3: **初始化：** 将 $\phi_j = 1$ 设为 $j = 1, \ldots, p$, 并将 $\boldsymbol{\beta}$设为普通最小二乘估计4: **重复**5: 给定当前的 ϕ_j，通过求解以下问题估计 $\boldsymbol{\beta}$：

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta}\|_2^2 + \lambda_0 \sum_{j=1}^{p} \phi_j (|\beta_j| + |\beta_{j+p}|) \right\}$$

6: 给定当前的 $\boldsymbol{\beta}$，通过最大化以下目标更新 $\boldsymbol{\lambda}$：

$$-n\lambda_0 \sum_{j=1}^{p} \exp\left(\sum_{l=1}^{L} \lambda_l A_{jl}/d \right) (|\beta_j| + |\beta_{j+p}|) - \frac{1}{2\tau^2} \sum_{l=1}^{L} \lambda_l^2$$

更新 $\phi_j = \exp\left(\sum_{l=1}^{L} \lambda_l A_{jl}/d\right)$，针对 $j = 1, \ldots, p$

7:

8: **直至** $\boldsymbol{\beta}$ 和 $\boldsymbol{\lambda}$收敛

9: **交叉验证：** 使用当前的 λ_0 和 ϕ_j计算交叉验证误差 c

10: **如果** $c < c^f$，**则**

11: 更新 $\lambda_0^f \leftarrow \lambda_0$、$c^f \leftarrow c$ 和 $\phi_j^f \leftarrow \phi_j$对 $j = 1$、$\ldots$、p

12: **结束如果**

13: **结束循环**

14: **最终拟合：** 使用 $\{\phi_j^f\}_{j=}^{p}$ 执行自适应 Lasso $_1$ 并 λ_0^f 以估计 $\boldsymbol{\beta}$

15: **输出：** $\boldsymbol{\beta}$、$\{\phi_j^f\}_{j=}^{p}$的最终估计值 $_1$ 以及 λ_0^f

2.4 利用汇总统计量的注释信息 Knockoffs

考虑一般情形：我们使用多元线性回归模型将响应向量 $\mathbf{y}$ 与设计矩阵 $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ 联系起来：

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{b}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n),$$

其中，$\mathbf{y}$ 和$\mathbf{X}$ 均已标准化为零均值和单位方差。向量$\boldsymbol{b}$ 表示这些 p 协变量的真实效应大小。

在实践中，由于隐私限制，大多数全基因组关联研究可能难以获取个体水平数据，其中包括表型向量$\mathbf{y}$ 和基因型矩阵 $\mathbf{X}$。因此，这些研究通常会提供汇总统计量，用于反映每个协变量的边际关联证据 j。

Algorithm 2 注释知情敲除（AnnoKn-lite）的简化版本

要求： 响应向量$\mathbf{y}$、协变量矩阵$\mathbf{X}$、敲除副本$\tilde{\mathbf{X}}$、注释矩阵$\mathbf{A}$、超参数 d, τ^2

初始化： 为 $j = 1, .., p$设置 $\phi_j = 1$，并在标准 Lasso 问题上通过交叉验证选择最优的 λ_0

重复

给定当前的 ϕ_j，通过求解以下问题估计 $\boldsymbol{\beta}$：

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta}\|_2^2 + \lambda_0 \sum_{j=1}^{p} \phi_j (|\beta_j| + |\beta_{j+p}|) \right\}$$

给定当前的 $\boldsymbol{\beta}$，通过最大化以下内容更新 $\boldsymbol{\lambda}$：

$$-n\lambda_0 \sum_{j=1}^{p} \exp\left(\sum_{l=1}^{L} \lambda_l A_{jl}/d \right) (|\beta_j| + |\beta_{j+p}|) - \frac{1}{2\tau^2} \sum_{l=1}^{L} \lambda_l^2$$

为 $j = 1, \ldots, p$更新 $\phi_j = \exp\left(\sum_{l=1}^{L} \lambda_l A_{jl}/d\right)$ **直到** $\boldsymbol{\beta}$ 和 $\boldsymbol{\lambda}$收敛 **重新调参：**

根据 $\{\phi_j\}_{j=1}^{p}$的最终估计值，通过交叉验证重新调整 λ_0

$\{\phi_j\}_{j=1}^{p}$ λ_0 $\boldsymbol{\beta}$ **最终拟合：**

应用自适应 Lasso，并使用 和 估计 **输出：** 的最终估计值、 $\{\phi_j\}_{j=1}^{p}$ 和 $\boldsymbol{\beta}$ λ_0

ˆ 汇总统计量包括估计的边际效应大小（b_j）及其对应的标准误（$se(b_j)$），通常通过针对每个协变量分别运行简单线性回归模型得到：

$$\mathbf{y} = \mathbf{x}_j b_j + \boldsymbol{\epsilon}_j, \quad \boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \sigma_j^2).$$

敲除框架已扩展为完全基于汇总统计量运行。具体而言，GhostKnockoff 方法 [29, 30] 仅利用边际 Z得分以及估计的协变量相关矩阵来执行受控变量选择；该相关矩阵可从外部数据源获取，例如千人基因组计划（1KG）的参考面板。

记$\hat{\boldsymbol{b}} = (\hat{b}_1, \ldots, \hat{b}_p)^\top$ 为通过边际线性回归估计得到的效应大小向量。在假设每个遗传变异与表型$\mathbf{y}$的相关性都很小的情况下， [31]中证明了

$$\hat{\boldsymbol{b}} | \boldsymbol{b}, \boldsymbol{S}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\boldsymbol{S}\boldsymbol{\Sigma}\boldsymbol{S}^{-1}\boldsymbol{b}, \boldsymbol{S}\boldsymbol{\Sigma}\boldsymbol{S}\right), \tag{5}$$

ˆ ˆ 其中 $\boldsymbol{S} = \text{diag}(se(b_1) \ldots, se(b_p))$，$\boldsymbol{\Sigma}$ 是连锁不平衡矩阵，表示 p单核苷酸多态性基因型之间的两两相关性。

将 $z_j = \hat{b}_j / se(\hat{b}_j)$ 记为单核苷酸多态性 j 的边际 Z-score，$\boldsymbol{z} = (z_1, \ldots, z_p)^\top$ 是边际 Z-score 向量。$\boldsymbol{z} = \boldsymbol{S}^{-1}\hat{\boldsymbol{b}}$ 的条件分布为

$$\boldsymbol{z}|\boldsymbol{b}, \boldsymbol{S}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\boldsymbol{\Sigma}\boldsymbol{S}^{-1}\boldsymbol{b}, \boldsymbol{\Sigma}\right). \tag{6}$$

对于多基因性状，任何单个单核苷酸多态性所解释的表型方差通常都可以忽略不计，因此我们可以假设 $\sigma_j^2 \approx 1$。此外，由于每个基因型向量 $\mathbf{x}j$ 均已标准化，我们有 $\mathbf{x}_j^\top \mathbf{x}j = n$。因此，$\hat{}\, b_j$ 的标准误可计算为

$$se(\hat{b}_j) = \sqrt{\sigma_j^2 / \mathbf{x}_j^\top \mathbf{x}_j} \approx \sqrt{1/n}.$$

采用这一近似后，式（5 和 6）中的条件分布可简化为

$$\hat{\boldsymbol{b}}|\boldsymbol{b}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\boldsymbol{\Sigma}\boldsymbol{b}, \boldsymbol{\Sigma}/n\right),$$

$$\boldsymbol{z}|\boldsymbol{b}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\sqrt{n}\boldsymbol{\Sigma}\boldsymbol{b}, \boldsymbol{\Sigma}\right).$$

给定 $\boldsymbol{z}$ 的这一条件分布，我们可以扩展 AnnoKn，使其仅利用汇总统计量以及估计的连锁不平衡矩阵 $\boldsymbol{\Sigma}$运行；该矩阵可从外部参考面板获得。我们采用多重 knockoff 框架，同时生成 M 个 knockoff 副本的 Z 分数（$\tilde{\boldsymbol{z}}_1$，……，$\tilde{\boldsymbol{z}}_M$）。按照 [32],，原始协变量与 M 个 knockoff 副本之间的联合相关结构形式如下：

$$\boldsymbol{\Sigma}_M = \underbrace{\begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}-\mathbf{D} & \ldots & \boldsymbol{\Sigma}-\mathbf{D} \\ \boldsymbol{\Sigma}-\mathbf{D} & \boldsymbol{\Sigma} & \ldots & \boldsymbol{\Sigma}-\mathbf{D} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}-\mathbf{D} & \boldsymbol{\Sigma}-\mathbf{D} & \ldots & \boldsymbol{\Sigma} \end{pmatrix}}_{M+1 \text{ blocks}},$$

其中，$\mathbf{D} = \text{diag}(s_1$，……，$s_p)$ 是一个对角矩阵，由求解以下凸优化问题得到：

$$\text{minimize} \quad \sum_j |1 - s_j|, \quad \text{subject to} \quad \begin{cases} \frac{M+1}{M}\boldsymbol{\Sigma} - \mathbf{D} \succeq 0, \\ s_j \geq 0, 1 \leq j \leq p. \end{cases}$$

将 $\boldsymbol{Z}_M = (\boldsymbol{z}^\top$，$\tilde{\boldsymbol{z}}_1^\top$、……、$\tilde{\boldsymbol{z}}_M^\top)^\top$ 记为一个包含原始协变量和 M knockoff 副本的联合边际 Z-分数的（$M+1$）p 向量，则 $\boldsymbol{Z}_M$ 的条件分布为：

$$\boldsymbol{Z}_M|\boldsymbol{\beta}, n, \boldsymbol{\Sigma}_M \sim \mathcal{N}\left(\sqrt{n}\boldsymbol{\Sigma}_M\boldsymbol{\beta}, \boldsymbol{\Sigma}_M\right),$$

其中，$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{(M+1)p})^\top$ 是原始协变量和M knockoff 副本的真实系数向量。

基于这一条件分布，我们在以下贝叶斯框架下，使用汇总统计量和估计的连锁不平衡矩阵，实施注释信息引导的 GhostKnock-off（AnnoGK）。与 AnnoKn 类似，我们也为系数设置拉普拉斯先验：

$$\boldsymbol{Z}_M|\boldsymbol{\beta}, n, \boldsymbol{\Sigma}_M \sim \mathcal{N}\left(\sqrt{n}\boldsymbol{\Sigma}_M\boldsymbol{\beta}, \boldsymbol{\Sigma}_M\right),$$

$$\beta_{j+k\times p} \sim Laplace(0, (n\lambda_0\phi_j)^{-1}), \quad k = 0, \ldots, M, \quad j = 1, \ldots, p,$$

$$\phi_j = \exp\left(\sum_{l=1}^{L} \lambda_l A_{jl}/d\right), \quad \lambda_l \sim \mathcal{N}(0, \tau^2),$$

其中 n 为样本量。然后，我们可以计算目标参数 $\boldsymbol{\beta}$ 和 $\boldsymbol{\lambda}$ 的后验分布：

$$\begin{aligned}
&p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{Z}_M, \boldsymbol{\Sigma}_M, n, \mathbf{A}, \boldsymbol{\theta}) \\
\propto &p(\boldsymbol{Z}_M|\boldsymbol{\beta}, \boldsymbol{\Sigma}_M, n, \boldsymbol{\theta}) \times p(\boldsymbol{\beta}|\mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\theta}) \times p(\boldsymbol{\lambda}|\boldsymbol{\theta}) \\
\propto &\exp\left(-\frac{1}{2}(\boldsymbol{Z}_M - \sqrt{n}\boldsymbol{\Sigma}_M\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_M^{-1}(\boldsymbol{Z}_M - \sqrt{n}\boldsymbol{\Sigma}_M\boldsymbol{\beta})\right) \\
&\times \prod_{j=1}^{p}\prod_{k=0}^{M}\left(\frac{n\lambda_0\phi_j}{2}\right)\exp(-n\lambda_0\phi_j|\beta_{j+k\times p}|) \times (2\pi\tau^2)^{-L/2}\exp\left(-\frac{1}{2\tau^2}\sum_{l=1}^{L}\lambda_l^2\right).
\end{aligned}$$

对数似然为

$$\begin{aligned}
&\log p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{Z}_M, \boldsymbol{\Sigma}_M, n, \mathbf{A}, \boldsymbol{\theta}) \\
=&-\frac{1}{2}(\boldsymbol{Z}_M - \sqrt{n}\boldsymbol{\Sigma}_M\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_M^{-1}(\boldsymbol{Z}_M - \sqrt{n}\boldsymbol{\Sigma}_M\boldsymbol{\beta}) + (M+1)\sum_{j=1}^{p}\log\phi_j \\
&-n\lambda_0\sum_{j=1}^{p}\phi_j\sum_{k=0}^{M}|\beta_{j+k\times p}| - \frac{1}{2\tau^2}\sum_{l=1}^{L}\lambda_l^2 + \text{const} \\
=&-\frac{n}{2}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_M\boldsymbol{\beta} + \sqrt{n}\boldsymbol{\beta}^\top\boldsymbol{Z}_M + (M+1)\sum_{j=1}^{p}\log\phi_j - n\lambda_0\sum_{j=1}^{p}\phi_j\sum_{k=0}^{M}|\beta_{j+k\times p}| - \frac{1}{2\tau^2}\sum_{l=1}^{L}\lambda_l^2 + \text{const}.
\end{aligned}$$

给定固定的超参数 λ_0，我们迭代更新 $\boldsymbol{\beta}$ 和 $\boldsymbol{\lambda}$，以最大化此对数后验：

- 给定 $\boldsymbol{\lambda}$，通过最小化损失函数来更新 $\boldsymbol{\beta}$：

$$\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_M\boldsymbol{\beta} - \frac{1}{\sqrt{n}}\boldsymbol{\beta}^\top\boldsymbol{Z}_M + \lambda_0\sum_{j=1}^{p}\phi_j\sum_{k=0}^{M}|\beta_{j+k\times p}|$$

使用 BASIL 框架 [33]。

- 给定 $\boldsymbol{\beta}$，通过最大化目标函数来更新 $\boldsymbol{\lambda}$：

$$
\begin{aligned}
&(M+1)\sum_{j=1}^{p}\phi_j - n\lambda_0\sum_{j=1}^{p}\phi_j(\sum_{k=0}^{M}|\beta_{j+k\times p}|) - \frac{1}{2\tau^2}\sum_{l=1}^{L}\lambda_l^2 \\
=&(M+1)\sum_{j=1}^{p}\sum_{l=1}^{L}\lambda_l A_{jl}/d - n\lambda_0\sum_{j=1}^{p}\exp(\sum_{l=1}^{L}\lambda_l A_{jl}/d)(\sum_{k=0}^{M}|\beta_{j+k\times p}|) - \frac{1}{2\tau^2}\sum_{l=1}^{L}\lambda_l^2 \\
=&-n\lambda_0\sum_{j=1}^{p}\exp(\sum_{l=1}^{L}\lambda_l A_{jl}/d)(\sum_{k=0}^{M}|\beta_{j+k\times p}|) - \frac{1}{2\tau^2}\sum_{l=1}^{L}\lambda_l^2. \qquad (7)
\end{aligned}
$$

我们在补充材料第 3.4 节总结了 AnnoGK 的算法。为选择最优值 λ_0，我们采用 [34] 提出的、并由 GhostKnockoff 在 [30] 中引用的伪汇总统计量方法。其核心思想是分别使用指定的训练样本量 n_t 和验证样本量 n_v，构建训练和验证汇总统计量。对于可能的 λ_0 值构成的候选集 $\mathbb{A}$，我们基于伪验证汇总统计，近似估计预测的 Lasso 系数与其真实对应值之间的相关性，并通过最大化该估计值来确定表现最佳的值。本研究采用五折交叉验证方案，其中 $n_t = 0.8n$ 和 $n_v = 0.2n$。此外，为了验证 AnnoGK 的基本用法，我们在全文中固定 $M = 1$。

3 个体水平数据的模拟研究

在本节中，我们使用个体水平数据**X**和**y** 开展模拟研究，以说明 AnnoKn 如何通过有效利用注释信息，在维持有效错误发现率控制的同时提高变量选择的检验效能。遵循 AdaKn 的模拟设置，我们考虑两种情形：一维和二维注释。

3.1 一维注释信息

我们首先复现了 AdaKn 研究中的主要模拟情景（详情见补充材料第 S4.1 节）。为模拟真实的基因型数据，协变量由隐马尔可夫模型（HMM）生成 [35]。模拟数据集包含 $n = 1{,}000$ 个样本和 $p = 900$ 个协变量。我们从前 300 个协变量中随机选取 150 个因果协变量，使得被选中的概率随协变量索引增大而降低。每个因果协变量均被赋予一个系数$\beta_j = \pm 3.5/\sqrt{n}$，而所有非因果协变量的系数均设为 0。注释定义为协变量索引：$A_j = j$ 对于$j = 1, \ldots, p$。

我们重复进行这项模拟100次，并将 AnnoKn 与 AdaPT、Knockoffs 以及采用多种敲除过滤器的 AdaKn 进行比较，包括广义线性模型（GLM）、广义加性模型（GAM）、随机森林（RF）、

以及双组模型（期望最大化）。如图1所示，AnnoKn 和 AnnoKn-lite 的表现始终优于所有 AdaKn 方法、AdaPT 和原始 Knockoff 程序（该程序忽略了注释信息）；在将假发现率（FDR）维持在目标水平以下的同时，检验效能显著更高。值得注意的是，尽管结构有所简化，AnnoKn-lite 的检验效能略低于 AnnoKn，但两者几乎相同。

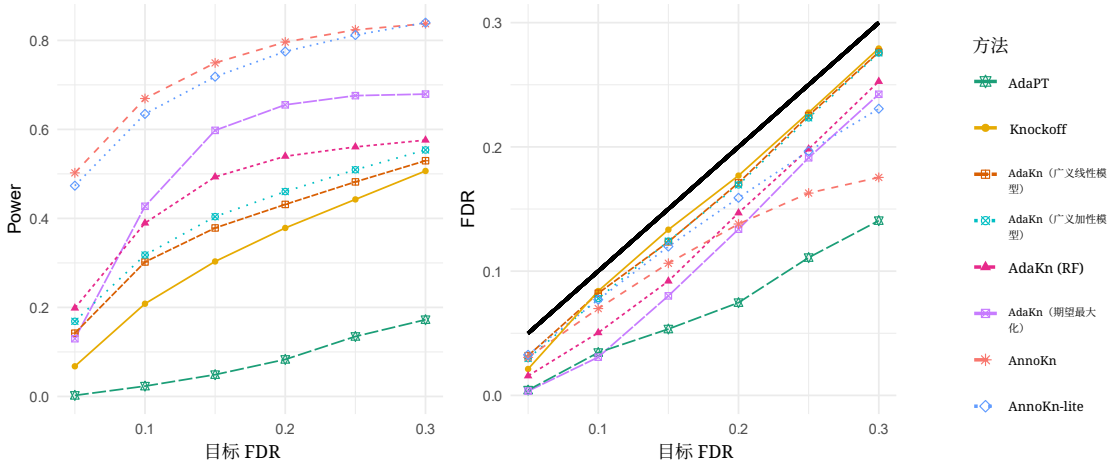

图1：**比较 Knockoffs、AnnoKn、AdaPT 和 AdaKn 在个体级数据和一维连续注释下的表现**。目标 q值范围为 0.05 到 0.3，其中黑色实线表示目标 q值水平。

我们还将连续指标转换为二元注释后进行了相同的模拟。具体而言，对于 $j \in \{1, \ldots, 300\}$，我们设为 $A_j = 0$，其他情况下则设为 $A_j = 1$（补充图 S1）。尽管其检验效能低于使用连续注释时的检验效能（这并不意外，因为二元注释信息包含的信息较少），AnnoKn 和 AnnoKn-lite 仍优于其他竞争方法。AdaPT 无法纳入二元注释，而 AdaKn (RF) 在严格的错误发现率控制下表现不佳。

3.2 二维注释信息

按照 AdaKn 的设置，我们模拟了包含 $n = 1{,}000$ 个样本和 $p = 1{,}600$ 个协变量的数据集。$n \times p$ 设计矩阵 $\mathbf{X}$ 由独立同分布的高斯分布元素生成。给定 $\mathbf{X}$，使用逻辑回归模型生成响应向量 $\mathbf{y}$（详情参见补充材料第 S4.2 节）。系数向量 $\boldsymbol{\beta}$ 在 40×40 网格上按空间结构排列。为了全面评估 AnnoKn 与其他方法的性能，我们考虑了四种系数可能具有因果效应的区域类型（补充图 S2）。在每个因果区域内，每个系数 β_j 是否非零的概率由其在平面上的对应坐标 $\big(r(j)\ \ s(j)\big)$ 决定；非零系数设为$25/\sqrt{n}$，其符号则由独立抛硬币决定。

—

我们首先将 AnnoKn 和 AnnoKn-lite 与 AdaKn 及 Knockoffs 进行比较，并将每个系数的坐标 $(r(j)\ \ s(j))$ 作为其二维连续注释。如图 2 所示，所有方法均能稳定地控制假发现率。然而，AnnoKn 和 AnnoKn-lite 的检验效能显著高于 Knockoffs 以及采用期望最大化和随机森林滤波器的 AdaKn 变体。AnnoKn 的检验效能与 AnnoKn-lite 相近，但假发现率更低。

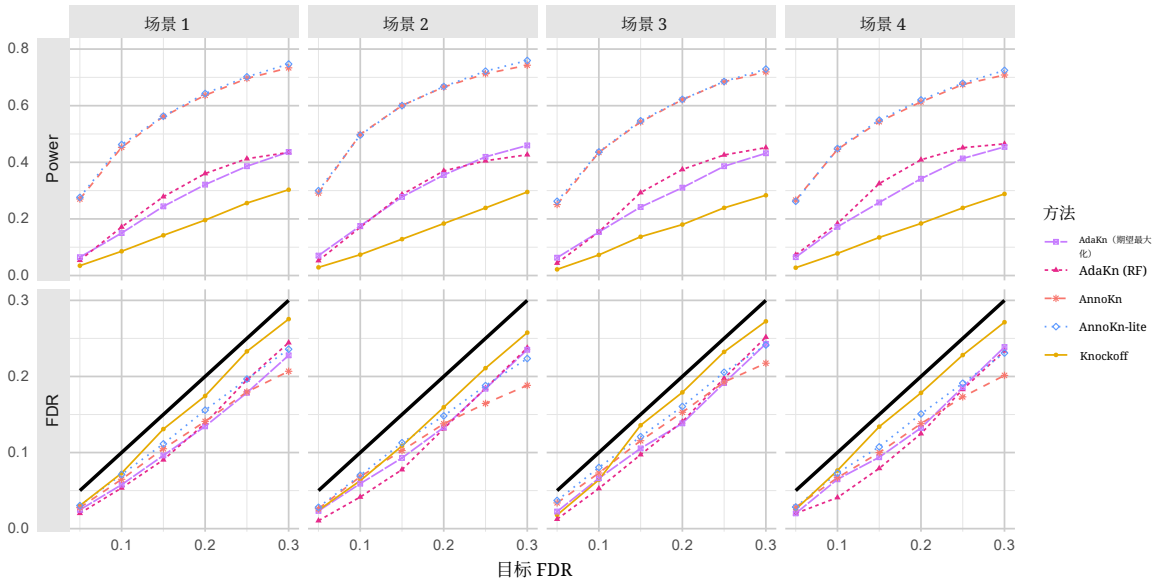

图 2：**基于个体水平数据和二维连续注释，对 Knockoffs、AnnoKn 和 AdaKn 进行比较。**目标 q值的范围为 0.05 至 0.3，黑色实线表示理论 q值水平。

接下来，我们评估了二元注释下的性能。具体而言，对于其 r坐标不超过因果区域最大 r 坐标的所有协变量，我们设定 $r(j) = 0$，否则设定 $r(j) = 1$。以类似方式将 $s(j)$坐标二值化。如补充图 S3所示，AnnoKn 和 AnnoKn-lite 均继续控制假发现率并提高统计功效。

为评估 AnnoKn 对非信息性注释的稳健性，我们通过随机置换 $\{1, \ldots, p\}$，在一维注释设置中额外加入了四个非信息性注释。如补充图 S4 和 S6 所示，并在补充材料第 S5.1 节中讨论，AnnoKn 和 AnnoKn-lite 相较于 AdaKn，在纳入噪声时表现出很强的稳健性。具体而言，AnnoKn 会自动降低非信息性注释的影响，将其对应参数 (λ_l) 向零收缩（补充图 S5 和 S7）。

4 基于汇总统计量的模拟研究

为说明通过 AnnoGK 将注释信息整合到 GhostKnockoff 中能够提升模型性能，我们在与 GhostKnockoff [30] 所用设置相似的条件下进行了数值模拟。我们考虑了三种样本量设置：$(n,\ p) \in \{(5,000,300), (5,000,600), (5,000,1,000)\}$。我们生成 $\mathbf{x} \overset{\text{iid}}{\sim} \mathcal{N}(0,\ \Sigma_\rho)$，用于 $i \in \{1, 2, ..., n\}$，其中 $[\Sigma_\rho]_{s,t} = \rho^{|s-t|} \quad 1 \le s,\ t \le p \qquad \rho = 0.5$。我们固定以在协变量之间引入中等程度的相关性。

响应变量生成方式为 $y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon_i$，其中 $\epsilon_i \overset{\text{独立同分布}}{\sim} \mathcal{N}(0,\ \sigma^2)$。我们通过从前 60 个坐标中选取 30 个设为非零，构造了一个稀疏系数向量 $\boldsymbol{\beta}$；其中第 j 个协变量为因果因素的概率与 j^{-2} 成正比。非零系数对应信号的符号以相同概率设为正或负；每个信号的绝对值相同，并通过参数 σ^2 控制遗传因素 h^2 对变异的解释比例：$= \text{Var}(\boldsymbol{\beta}^\top \mathbf{x}) / \text{Var}(\mathbf{y})$ $\in \{0.05, 0.1, 0.2\}$。

我们首先考虑将每个协变量的索引作为连续注释信息。对于每种设置，我们考察了四种方法：Knockoffs（使用 LCD统计量）、AnnoKn、GhostKnockoff（使用 LCD统计量，调参参数 λ_0 通过伪汇总统计方法选取）以及 AnnoGK。图 3 展示了当 $p = 1,000$ 时，各方法的检验效能和假发现率随目标 FDR 水平的变化情况。在所有遗传力取值下，各方法均能恰当地控制假发现率。与各自对应的方法 GhostKnockoff 和 Knockoffs 相比，注释信息方法（AnnoGK 和 AnnoKn）始终取得更高的检验效能和更低的实际假发现率。基于汇总统计量的方法往往更保守，对假发现率的控制更严格；而 AnnoGK 的检验效能与 AnnoKn 非常接近，尤其是在信号较强时（$h^2 = 0.2$）。如补充图S9和S10所示，这一模式在 $p \in \{300, 600\}$ 中也成立：在不同目标 FDR 水平下，检验效能和假发现率的变化趋势几乎相同。这些结果表明，相较于 GhostKnockoff，AnnoGK 在提升检验效能的同时，能够更严格地控制假发现率，表现稳健。

我们还使用每个协变量的二元注释评估了所有方法的性能：当索引小于或等于 60 时，注释设为 1；否则设为 0。如补充图S11、S12和S13所示，无论注释类型如何，AnnoKn 和 AnnoGK 均表现稳健。与 Knockoffs 和 GhostKnockoff 相比，在二元和连续注释下，AnnoKn 和 AnnoGK 均持续提升检验效能并降低假发现率，且不同注释类型下的提升幅度相近。

为使用真实基因型数据验证 AnnoKn 和 AnnoGK，我们利用 10,000 名 UK Biobank（UKBB）参与者进行了模拟。连锁不平衡矩阵使用样本内数据或外部参考面板进行估计。如补充图 S14 和 S15 所示（并在补充材料第 5.2 节中讨论），使用样本内连锁不平衡矩阵时，AnnoKn 和 AnnoGK 提高了检测因果单核苷酸多态性的检验效能，同时降低了假发现率。此外，当我们使用外部参考面板估计

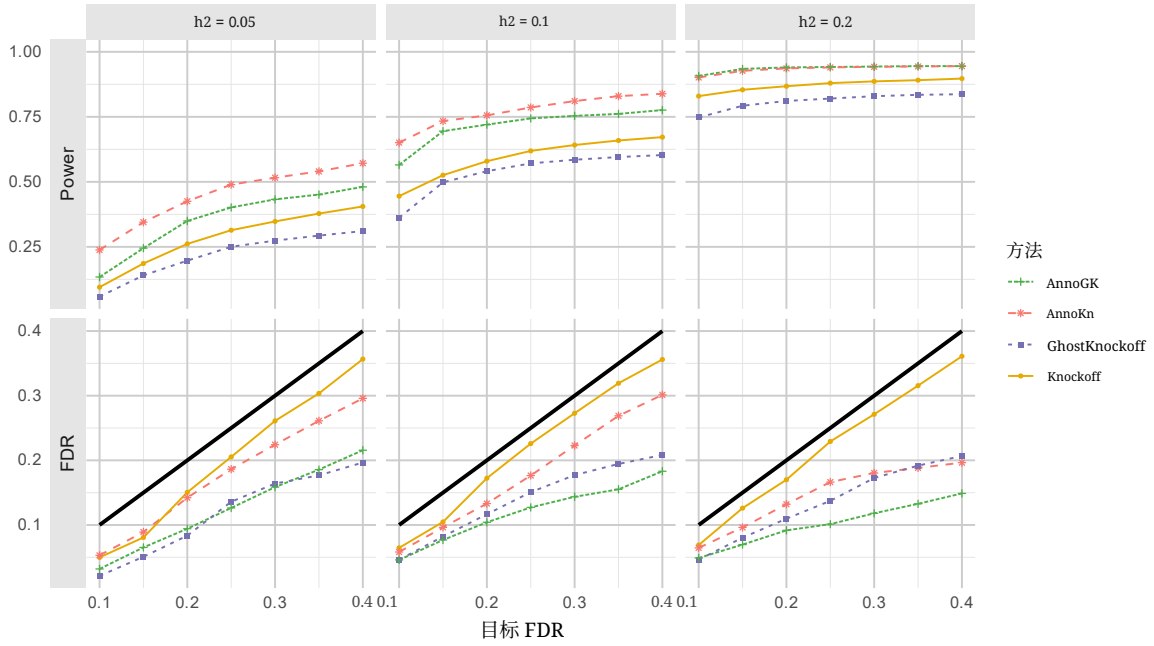

图 3：**使用模拟的 AR(1)设计矩阵，对比 Knockoffs、AnnoKn、GhostKnockoff 和 AnnoGK**。我们在 $\{0.05, 0.1, 0.2\}$ 范围内改变由协变量解释的表型方差（h^2）。 q值阈值介于 0.1 至 0.4 之间，黑色实线表示理论上的 q值水平。协变量数量为 $p = 1,000$。我们使用每个协变量的索引作为连续注释信息。

在连锁不平衡矩阵方面，与 GhostKnockoff 相比，AnnoGK 对参考面板与汇总统计量之间的 LD 不匹配表现出更强的稳健性，减轻了这种不匹配导致的 FDR 膨胀。

最后，我们评估了 AnnoKn 和 AnnoGK 的计算效率。我们将 AnnoKn 和 AnnoKn-lite 的运行时间与 AdaPT、Knockoffs 和 AdaKn 进行了比较（补充图S8）。在一维注释设置下，AnnoKn 的计算效率与 AdaKn 相当，且随着注释维度增加，其效率优势也逐渐显现。此外，AnnoKn-lite 无需在候选集中搜索最优的 λ_0，因此与 AnnoKn 相比，计算时间大幅缩短。由此，AnnoKn-lite 成为效率最高的注释信息引导的变量选择方法之一。此外，将 Knockoffs 和 AnnoKn 替换为相应的基于汇总统计量的版本，进一步缩短了计算时间（补充图S16），突显了在大规模关联研究中使用汇总统计量相较于个体水平数据的优势。

5 真实数据应用

5.1 AnnoKn改进了GTEx表达数量性状基因的识别

我们应用AnnoKn在基因型-组织表达（GTEx）项目v8 [36]中检测表达数量性状基因（其表达水平至少受一个独立表达数量性状位点（eQTL）影响的基因）。预处理的详细信息见补充材料S7.5和S7.6节。GTEx数据集包含组织特异性基因表达测量值和全基因组测序数据，各组织的样本量汇总于补充图S17。本研究将分析范围限定为样本量最大的十种GTEx组织中的蛋白质编码基因，并排除了顺式SNP数量不超过10个的基因。具有可用表达数据的基因数量汇总于补充表S1。对于每个基因，我们纳入位于基因内部以及其上游和下游250 kb范围内的顺式SNP。

与标准 Knockoff 应用一致，我们基于连锁不平衡对单核苷酸多态性进行层次聚类，使用的绝对两两相关性阈值为 $|r| > 0.5$。随后，我们从每个簇中选取一个代表性SNP，具体来说，是与表型边际关联最强的单核苷酸多态性（其 p 值最小）。

我们从分层LD评分回归（S-LDSC） [37, 38], 中获取功能注释信息，该方法为每个单核苷酸多态性提供 96 维注释。我们首先使用三种代表性注释比较 AnnoKn 与 Knockoffs 的表现，这些注释与表达数量性状位点的关联已得到充分研究：启动子 [39, 40], 增强子 [36, 41], 和转录因子结合位点（TFBS） [42, 43]。在 q值阈值为 0.1 或 0.2 时，在所有评估的组织中，AnnoKn 识别出的表达数量性状基因均多于 Knockoffs，平均增幅分别为 12.8% 和 8.1%（表1和补充图S18）。

表1：在 q-值阈值为 0.1 或 0.2 时，使用三种具有代表性的 S-LDSC 注释，在十种组织中由 Knockoffs 和 AnnoKn 识别出的表达数量性状基因数量。

组织	Knockoffs		AnnoKn	
	$\mathbf{q = 0.1}$	$\mathbf{q = 0.2}$	$\mathbf{q = 0.1}$	$\mathbf{q = 0.2}$
骨骼肌	94	1,178	110	1,292
全血	75	1,229	81	1,322
日光暴露皮肤（小腿）	96	1,409	102	1,519
胫动脉	115	1,424	138	1,524
皮下脂肪	92	1,336	106	1,473
甲状腺	119	1,568	143	1,661
胫神经	149	1,577	170	1,688
未暴露于阳光的皮肤（耻骨上区）	99	1,420	110	1,556
Lung	104	1,266	112	1,387
食管黏膜	107	1,440	116	1,532

这些结果表明，AnnoKn 能够利用功能注释信息对 eQTL 进行优先级排序，从而识别出更多表达数量性状基因。我们进一步评估了 AnnoKn 识别出的表达数量性状基因在生物学上是否更具意义，以及是否在组织相关基因集中富集程度更高。作为外部验证资源，我们使用了 Jensen TISSUES 数据库 [44], ，这是一个组织表达数据库，整合了人工整理的文献、蛋白质组学、转录组筛查和自动化文本挖掘所得的证据。该数据库为组织-基因关联赋予置信度评分，从而能够对不同证据来源进行系统比较。对于每种 GTEx 组织，我们在 Jensen TISSUES 中选取了有充分证据支持的对应组织或最密切相关的组织（补充表 S1）。

我们评估了与所有候选基因相比，Knockoffs 和 AnnoKn 识别出的表达数量性状基因是否富集于组织特异性的 Jensen基因集。具体而言，我们使用 Enrichr [45, 46], 进行了富集分析，并将每种组织的所有候选基因定义为背景。由此，我们计算出一个 p 值，用以表示识别出的表达数量性状基因在匹配的 Jensen组织中富集的程度，从而定量衡量每种方法优先筛选组织相关基因的效果。

如图4所示，在10种GTEx组织中的9种组织里，AnnoKn识别出的表达数量性状基因在Jensen数据库中表现出更强的富集。例外是胫动脉，在该组织中，两种方法均未产生显著富集的表达数量性状基因。这可能是因为Jensen数据库缺少直接对应的组织；最接近的可用替代组织——冠状动脉——在生物学特性和表达数量性状位点基因谱方面存在差异。在显著性阈值为 $p < 0.05$时，除胫动脉外，AnnoKn在10种组织中的9种组织里实现了显著富集。相比之下，Knockoffs仅在10种组织中的5种组织里实现了显著富集。这些结果

表明，与标准敲除框架相比，AnnoKn通过纳入功能注释来优先筛选有信息量的eQTL，能够在多种组织中捕获更多且生物学意义更强的表达数量性状基因。

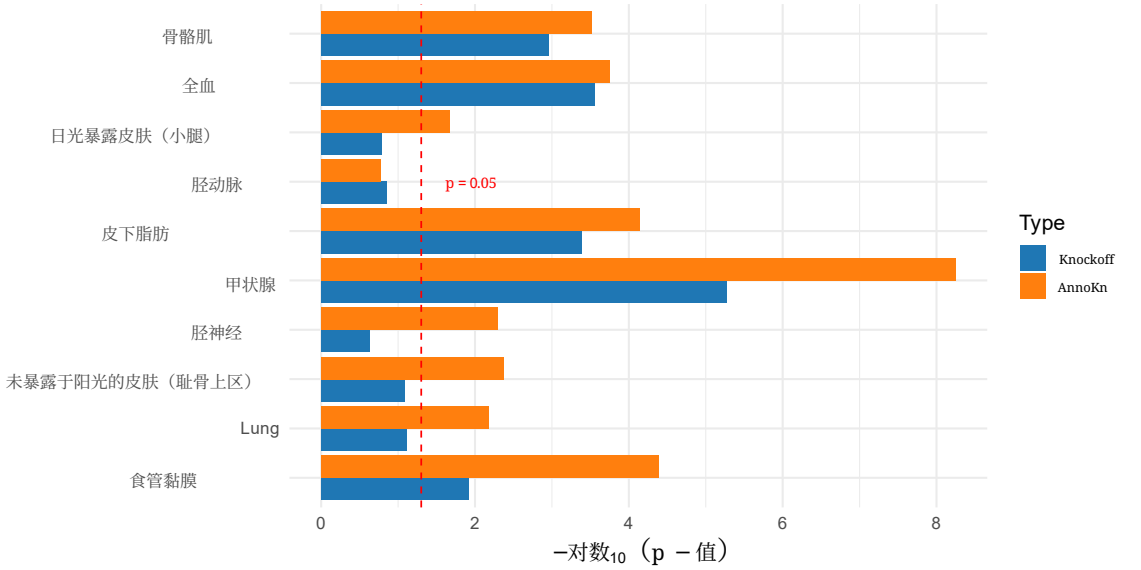

图4：**在 q值阈值为 0.2 时，比较 Knockoffs 和 AnnoKn 在 GTEx组织中识别出的表达数量性状基因的富集情况**。我们计算了在匹配的 Jensen组织中，已识别的组织特异性表达数量性状基因过度代表性的 –log(p值)。

我们还评估了每种功能注释对应的估计参数 λ_l 。对于启动子、增强子和 TFBS，这些注释均为二元变量：SNP 位于相应功能区域内时取值为 1，否则为 0。为将功能区域内的 SNP 优先列为潜在 eQTL，我们希望为其分配较小的自适应 Lasso 惩罚项，这对应于 λ_l的负值。为评估 AnnoKn 是否通过学习合理的参数值，有效地优先考虑功能性遗传变异，我们考察了每种注释下参数估计值显著为负（$\lambda_l < -1$）或显著为正（$\lambda_l > 1$）的基因数量。如补充图S19所示，显著参数的数量相较于蛋白质编码基因总数（约 18,000 个）较少（仅数百个），这表明对于大多数基因而言，功能注释对 eQTL 鉴定的影响总体较弱。在这些显著的参数值中，大多数为负值。在 GTEx组织中，增强子、启动子和 TFBS 对应的参数显著为负（$\lambda_l < -1$）的基因平均数量分别为 480.0、270.0 和 277.9。相比之下，这些注释对应的参数显著为正（$\lambda_l > 1$）的基因平均数量要少得多，分别为 13.1、15.0 和 18.2。这些结果表明，AnnoKn 成功利用了

功能注释，为生物学上相关的协变量分配较小的惩罚权重，从而提高 eQTL 的优先级排序效率并促进表达数量性状基因的识别。

为进一步展示 AnnoKn 对注释选择的稳健性，我们使用 S-LDSC 数据集中的全部 96 项功能注释来识别表达数量性状基因。如补充表 S2 和补充图 S20 所示（详见补充材料第 S6.1 节），纳入完整的注释集后，AnnoKn 在两个 q-值阈值下识别出的表达数量性状基因均多于 Knockoffs（阈值为 0.1 时多 13.7%，阈值为 0.2 时多 10.6%）。这些增益超过了仅使用三项代表性注释时的增益，不过额外提升相对有限。这种有限的提升或许是因为某些额外注释的相关性较低，且包含高度冗余的信息。此外，AnnoKn 识别出的表达数量性状基因比 Knockoffs 检出的基因表现出更强的组织特异性富集（补充图 S21）。综合来看，这些发现表明，最初选定的代表性注释已捕获了足够的功能信息，而纳入全部注释还能带来额外收益，但提升幅度相对有限。

5.2 AnnoGKimproves 跨祖源全基因组关联研究

为说明如何利用注释信息为 GhostKnockoff 提供信息，我们在两个旨在分别识别身高和精神分裂症（SCZ）遗传位点的全基因组关联研究示例中应用 AnnoGK，以利用跨祖源信息。

在第一项研究中，我们聚焦于一项涉及欧洲和非洲参与者的身高荟萃分析，并整合了 GIANT 联盟内多个队列的数据[47] （详情见补充材料第 S7.3 节）。补充图 S22 展示了两个祖源各自的曼哈顿图。对于这两个数据集，我们仅保留与 1KG 第 3 阶段参考面板 [48]中的单核苷酸多态性重叠的单核苷酸多态性。我们使用 Popcorn[49] 估算跨族群遗传相关性。如表 2 所示，欧洲与非洲祖源之间的身高遗传相关性为 0.70（标准误 $= 0.10$， p-值$= 0.002$），这有力地表明，与身高相关的大部分遗传变异在这两个祖源之间是共有的。

表 2：遗传力 身高和 SCZ 全基因组关联研究的遗传力和跨祖源遗传相关性 .

性状	祖源	遗传力（标准误）	遗传相关性（标准误）
身高	欧洲	0.38 (0.03)	0.70 (0.10)
	非洲	0.33 (0.03)	
精神分裂症	欧洲	0.60 (0.06)	0.75 (0.07)
	东亚人	0.62 (0.05)	

我们在欧洲人群身高全基因组关联研究中，以显著遗传关联信号（p-值 $< 5 \times 10^{-8}$）为中心界定区域，并将每个信号上游和下游 250 Kb 范围内的所有单核苷酸多态性纳入同一候选区域。我们合并相互重叠的区域，直至不再存在重叠。该流程在所有染色体上共得到 1,047 个区域（补充图 S23）。与 GTEx 分析一样，我们进行了层次聚类，将两两绝对相关性超过 0.5 的单核苷酸多态性归为一组，并选择每个簇中边际 p-值最小的单核苷酸多态性作为代表性SNP。

随后，我们使用 1KG 参考面板，将 GhostKnockoff 和 AnnoGK 应用于欧洲人群和非洲人群的全基因组关联研究数据，以识别这些区域内的推定因果单核苷酸多态性。我们将包含至少一个显著单核苷酸多态性的区域定义为显著区域。对于每种祖源，我们将另一种祖源中每个单核苷酸多态性的 –log(p-值) 作为连续注释，注释值越大，表示其与另一人群身高关联的证据越强。

在 0.1 的错误发现率阈值下，GhostKnockoff 在欧洲人群全基因组关联研究中，从 297 个区域识别出 8,737 个显著单核苷酸多态性；在非洲人群全基因组关联研究中，则从 132 个区域识别出 3,376 个显著单核苷酸多态性。纳入注释信息后，AnnoGK 在两种祖源中均大幅增加了显著单核苷酸多态性和显著区域的数量（表 3）。我们还比较了在两种祖源中均被识别出的区域数量，发现使用 AnnoGK 后增加了 44%。这一提升超过了任一单一祖源中的增幅，表明利用跨祖源遗传架构不仅能增加发现数量，还能提高所检测到的全基因组关联研究位点的可重复性。

表 3：GhostKnockoff 和 AnnoGK 在身高全基因组关联研究中识别出的因果单核苷酸多态性数量（错误发现率阈值 $= 0.1$）以及含有显著单核苷酸多态性的区域。

方法	**欧洲人群全基因组关联研究**		**非洲人群全基因组关联研究**		**共有区域** **在两项全基因组关联研究之间**
	区域	**SNP**	**区域**	**SNP**	
GhostKnockoff	297	8,737	132	3,376	88
AnnoGK	416	14,443	158	4,217	127
增加	40%	65%	20%	25%	44%

与 GTEx 分析类似，我们评估了 AnnoGK 在欧洲祖源全基因组关联研究中识别出的 416 个显著区域的估计注释参数 λ_l。我们预期，在非洲血统中，p 值较小的单核苷酸多态性在自适应 Lasso 模型中受到的惩罚较轻，因此 λ_l 应为负值，以优先考虑具有较大注释值（–log(p 值））的单核苷酸多态性。如补充图 S24 所示，估计的 λ_l 中有 78.8%（328/416）为负值，这表明 AnnoGK 有效地学到了预期模式，即优先考虑在另一个祖源中具有较强边际关联的单核苷酸多态性。

最后，我们比较了 GhostKnockoff 和 AnnoGK 在欧洲和非洲身高全基因组关联研究中识别出的显著区域的重叠情况。如图5所示，GhostKnockoff 检测到的大多数

区域也能被 AnnoGK 找到。然而，GhostKnockoff 识别出的区域并非都能被 AnnoGK 识别。这种差异可能部分由两个因素造成。首先，敲除构造固有的随机性会导致不同运行间的发现结果存在差异。其次，如果功能注释提供的支持不足，AnnoGK 可能无法检测到某个区域，从而使相应信号在注释信息引导的流程中被降低优先级。

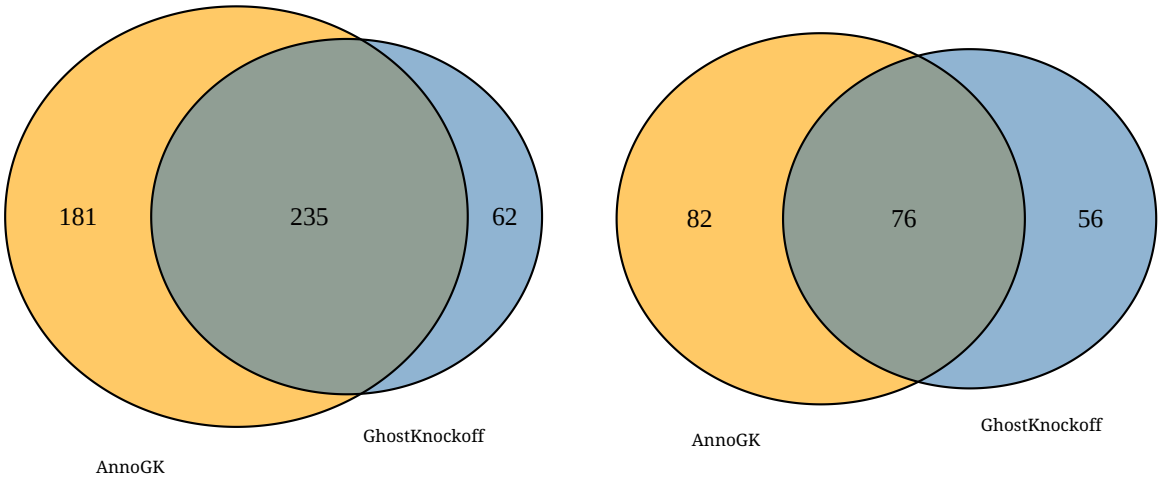

图5：**GhostKnockoff 和 AnnoGK 在欧洲祖源（左）和非洲裔美国人祖源（右）身高全基因组关联研究中检测到的显著区域维恩图。**

在第二个示例中，我们将 AnnoGK 应用于一项涉及四种祖源个体的精神分裂症荟萃分析：欧洲祖源（EUR）、东亚祖源（EAS）、非洲裔美国人祖源（AFR）和拉丁裔祖源（LAT） [50] （详情请参见补充图S25和补充材料第S7.4节）。在此分析中，我们主要关注样本量相对较大的两种祖源：EUR 和 EAS，同时将其余祖源的汇总统计量用作注释信息。如表 2 所示，EUR 与 EAS 之间的精神分裂症遗传相关性非常显著（ p- 值 $= 0.0003$ ），这表明借助不同祖源间的信息来检测精神分裂症的因果单核苷酸多态性具有重要意义。

按照身高全基因组关联研究应用中的相同步骤，我们基于欧洲祖源精神分裂症全基因组关联研究的汇总统计量获得了 169 个区域（补充图S26）。其中一个区域位于主要组织相容性复合体（MHC）区域，鉴于其连锁不平衡结构复杂，我们将其移除；这也是全基因组关联研究及其他下游分析中的常见做法 [51]。随后，我们以千人基因组数据集为参考面板，比较了 AnnoGK 与 GhostKnockoff 在检测包含显著因果单核苷酸多态性的区域方面的表现，分析对象为欧洲祖源和东亚祖源的精神分裂症全基因组关联研究。对于每种祖源，我们使用了来自其他三个祖源的 $-\log(p\text{-值})$

作为多维连续注释。

考虑到 knockoff程序 [52], 固有的随机性，重复运行时选出的协变量集合可能有所不同。对于精神分裂症全基因组关联研究，这种变异预计会更大，因为其关联总体上弱于身高研究。为缓解这一问题，我们使用 10 个独立的 knockoff 副本重复了这项分析。汇总所有重复的结果后，通过 AnnoGK 纳入其他祖源的信息，使欧洲祖源和东亚祖源研究中在错误发现率阈值为 0.1 时包含显著单核苷酸多态性的区域总数分别增加了 33% 和 23%（表 4）。此外，注释信息指导的变量选择使两个祖源中识别出的区域数从 2 个增加到 12 个。换言之，使用其他祖源的边际关联对候选单核苷酸多态性进行优先排序后，东亚祖源全基因组关联研究中 75%（12/16）的显著区域也在欧洲祖源全基因组关联研究中重现。

表 4：在欧洲祖源和东亚祖源的精神分裂症全基因组关联研究中，GhostKnockoff 和 AnnoGK 识别出的因果单核苷酸多态性数量（错误发现率阈值为 $= 0.1$）及包含显著单核苷酸多态性的区域数；结果汇总了 10 次 knockoff重复。

方法	欧洲祖源全基因组关联研究		东亚祖源全基因组关联研究		共同区域 两个全基因组关联研究之间
	区域	SNP	区域	SNP	
GhostKnockoff	40	907	13	198	2
AnnoGK	53	1,243	16	317	12
增加	33%	37%	23%	60%	500%

为了评估不同 knockoff重复之间的可重复性，我们考察了至少在给定比例的重复中被识别出的区域数量。如补充图 S27 所示，在所有检测频率阈值下，AnnoGK 识别出的区域数量均高于 GhostKnockoff。例如，当检测频率阈值至少为 0.5 时，AnnoGK 在欧洲祖源和东亚祖源全基因组关联研究中分别识别出 6 个和 1 个可重复区域，而 GhostKnockoff 仅识别出 2 个和 0 个。补充图S28 显示，可重复SNP的数量也有类似提升。这些结果表明，AnnoGK 利用注释信息更好地优先筛选因果单核苷酸多态性，从而提高了基于knockoff的变量选择的稳定性。

6 讨论

本文介绍了 AnnoKn，这是一种将功能注释纳入 knockoff程序的新型框架。该方法通过贝叶斯框架将自适应 Lasso 与 Knockoffs 相结合，从而区别于现有方法；它既能实现受控变量选择，又能在统一流程中利用功能注释对协变量进行优先级排序。我们进一步将 AnnoKn 扩展为 AnnoGK，后者使用

汇总统计量和参考面板来估计遗传相关性。这一扩展显著提升了注释知情敲除在大规模基因组学应用中的实用价值。

通过模拟和真实数据分析，我们展示了 AnnoKn 利用这些协变量的先验信息提升变量选择检验效能的优势。我们重点介绍了 AnnoKn 的两项具有实际意义且影响深远的应用。首先，AnnoKn 可以利用功能注释或多组学数据筛选优先级较高的因果遗传变异，如我们的 GTEx eGene 分析所示。其次，AnnoKn 提供了一种有原则的方法，可在不同祖源之间借用信息，从而改进代表性不足人群的关联研究；在这些人群中，样本量有限往往会降低统计功效。

AnnoKn 和 AnnoGK 的一项关键优势在于，它们能够高效处理高维注释数据，并自适应地识别对模型贡献最大的注释。这一能力免去了主观指定注释权重的需要，也无需在主要分析之前单独进行注释筛选预处理。另一项优势是其目标函数具有良好的性质，从而确保 AnnoKn 的稳定性和计算效率。在所有模拟场景和真实数据应用中，我们均未观察到 AnnoKn 和 AnnoGK 算法出现故障；相比之下，AdaKn 偶尔会遇到收敛问题，无法产生有效结果。此外，尽管由于生成 knockoff 副本，knockoff 程序本质上具有随机性，但 AnnoKn 有望利用注释作为一致的信息来帮助筛选优先级较高的候选协变量，从而提高 knockoff 方法的稳定性。

AnnoKn 灵活的建模框架也为未来的扩展和方法学发展开辟了多个方向。首先，当前模型将自适应 Lasso 的惩罚项与注释的线性组合相联系，因此该方法只能捕捉注释数据中的线性关系。鉴于大多数常用注释都有明确的方向性解释（例如，表示功能区域的指示变量，或代表更强关联的较小 p-值），这一假设通常已足够。不过，探索更灵活的连接函数，以适应现代生物医学及其他研究领域中日益丰富复杂的注释资源，仍然值得一试。

其次，利用 AnnoKn 借助跨祖源 GWAS 信息，可能看起来与其他荟萃分析变量选择方法相似，尤其是跨祖源精细定位方法 [53, 54, 55]。然而，出于以下几方面原因，我们发现很难将 AnnoKn 的结果与这些方法直接比较。(1) 大多数跨祖源精细定位方法采用贝叶斯方法，并以不同方式定义发现，重点是识别一组高度相关的单核苷酸多态性，其累积后验概率超过预先设定的阈值。相比之下，AnnoKn 会先将高度相关的协变量聚类为若干组，再以各组的代表变量进行错误发现率控制的变量选择。因此，其输出无法与后验纳入概率（PIPs）直接比较。(2) 尽管 AnnoKn 借鉴了多个祖源的信息，但其主要目标是利用外部信息改善主要研究中的推断。这与跨祖源精细定位不同，后者旨在发现共享和祖源特异的遗传变异。开发用于将 AnnoKn 与跨祖源精细定位方法进行比较的系统性方法仍是未来研究中一个有趣的方向，

采用系统化方法比较 AnnoKn 与跨祖源精细定位方法，仍是未来研究的一个有趣方向。

总之，AnnoKn 提供了一个可靠且有原则的框架，可将关于协变量的先验知识纳入高维非线性模型的变量选择中。这种方法解决了一项根本性的统计挑战，并广泛适用于统计遗传学以外众多科学领域。

7 致谢

本研究部分由美国国立卫生研究院资助，项目编号为 U24 HG012108 和 U01 HG013840。

8 利益声明

作者声明不存在利益冲突。

9 网络资源

我们使用从资源网站 https://www.ukbiobank. ac.uk/ 获取的 UKBB 基因型数据。GTEx数据集可从 https://www.ncbi.nlm.nih.gov/projects/ gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2 获取。1KG 参考面板可从 https://www.cog-genomics.org/plink/2.0/resources 获取。身高全基因组关联研究的汇总统计量可在 GIANT 联盟网站上找到（https://giant-consortium.web. broadinstitute.org/index.php/GIANT_consortium_data_files）。精神分裂症的汇总统计量由精神病学基因组学联盟（PGC）提供，网址为 https://figshare. com/articles/dataset/scz2022/19426775。Jensen组织表达数据库可从 http://tissues.jensenlab.org/ 获取。

参考文献

[1] Barbara E Stranger、Eli A Stahl 和 Towfique Raj。《人类复杂性状遗传学全基因组关联研究的进展与前景》。*Genetics*，187(2)：367–383，2011年。

[2]Eric R Gamazon、Heather E Wheeler、Kaanan P Shah、Sahar V Mozaffari、Keston Aquino-Michaels、Robert J Carroll、Anne E Eyler、Joshua C Denny、GTEx 联盟、Dan L Nicolae 等。用于利用参考转录组数据进行性状定位的基因关联分析方法。*Nat. Genet.*，47(9):1091–1098，2015年。https://doi.org/10.1038/ ng.3367.

[3] Alexander Gusev、Arthur Ko、Huwenbo Shi、Gaurav Bhatia、Wonil Chung、Brenda WJH Penninx、Rick Jansen、Eco JC De Geus、Dorret I Boomsma、Fred A Wright 等。大规模全转录组关联研究的整合方法。*Nat. Genet.*，48(3):245–252，2016年。https://doi.org/10.1038/ng.3506.[4] Peter M Visscher、Matthew A Brown、Mark I McCarthy 和 Jian Yang。全基因组关联研究发现的五年历程。《美国人类遗传学杂志》，90(1):7–24，2012年。[5]Peter M Visscher、Naomi R Wray、Qian Zhang、Pamela Sklar、Mark I McCarthy、Matthew A Brown 和 Jian Yang。全基因组关联研究发现十年：生物学、功能与转化。《美国人类遗传学杂志》，101(1):5–22，2017年。[6] Gregory M Cooper 和 Jay Shendure。在针堆中找针：从海量基因组数据中寻找疾病因果变异。《自然综述：遗传学》，12(9):628–640，2011年。[7] Eddie Cano-Gamez 和 Gosia Trynka。从全基因组关联研究到功能：利用功能基因组学识别复杂疾病背后的机制。《遗传学前沿》，11:424，2020年。[8] Qiongshi Lu、Xinwei Yao、Yiming Hu 和 Hongyu Zhao。GenoWAP：通过整合分析基因组功能注释对全基因组关联研究信号进行优先排序。*Bioinformatics*，32(4):542–548，2016年。[9] Daniel J Schaid、Wenan Chen 和 Nicholas B Larson。从全基因组关联到候选因果变异：统计精细定位。《自然综述：遗传学》，19(8):491–504，2018年。[10] Joseph K Pickrell、John C Marioni、Athma A Pai、Jacob F Degner、Barbara E Engelhardt、Everlyne Nkadori、Jean-Baptiste Veyrieras、Matthew Stephens、Yoav Gilad 和 Jonathan K Pritchard。利用 RNA 测序理解人类基因表达变异背后的机制。《自然》，464(7289):768–772，2010年。[11] GTEx联盟。人体组织中基因表达的遗传效应。《自然》，550(7675):204–213，2017年。[12] Kyoko Watanabe、Erdogan Taskesen、Arjen Van Bochoven 和 Danielle Posthuma。利用 FUMA 对遗传关联进行功能定位与注释。*Nature communications*，8(1):1826，2017年。[13] Xiaoquan Wen、Francesca Luca 和 Roger Pique-Regi。eQTL 的跨人群联合分析：精细定位与功能注释。*PLoS genetics*，11(4):e1005176，2015年。[14] Rekha Mudappathi、Tatiana Patton、Hai Chen、Ping Yang、Zhifu Sun、Panwen Wang、Chang-Xin Shi、Junwen Wang 和 Li Liu。reg-eQTL：整合转录因子效应以揭示调控变异。《美国人类遗传学杂志》，112(3):659–674，2025年。

[15]Qiongshi Lu、Chentian Jin、Jiehuan Sun、Russell Bowler、Katerina Kechris、Naftali Kaminski 和 Hongyu Zhao。通过数据整合进行GWAS后优先级排序，为慢性阻塞性肺疾病带来了新见解。*Statistics in biosciences*，9(2):605–621，2017年。[16] Yoav Benjamini 和 Yosef Hochberg。控制假发现率：一种实用且强大的多重检验方法。*Journal of the Royal statistical society: seriesB (Methodological)*，57(1):289–300，1995。[17] Yoav Benjamini 和 Daniel Yekutieli。依赖情形下多重检验中假发现率的控制。*Annals of statistics*，第1165–1188页，2001年。[18] Lihua Lei 和 William Fithian。AdaPT：一种利用侧信息进行多重检验的交互式程序。*Journal of the Royal Statistical Society Series B: Statistical Methodol-ogy*，80(4):649–679，2018年。[19] Yoav Benjamini。发现假发现率。*Journal of the Royal StatisticalSociety Series B: Statistical Methodology*，72(4):405–416，2010年。[20] Damian Brzyski、Christine B Peterson、Piotr Sobczyk、Emmanuel J Cand`es、Malgorzata Bogdan 和 Chiara Sabatti。控制GWAS假发现率。*Genetics*，205(1):61–75，2017年。[21] Jianliang He、Bowen Gang 和 Luella Fu。多重检验中的假发现控制：理论与方法简述。*arXiv preprint arXiv:2411.10647*，2024。[22] Jeffrey C Barrett、B Fry、JDMJ Maller 和 Mark J Daly。Haploview：连锁不平衡和单倍型图谱的分析与可视化。*Bioinformatics*，21(2):263–265，2005。[23] Rina Foygel Barber 和 Emmanuel J. Cand`es。通过Knockoff控制假发现率。*The Annals of Statistics*，43(5):2055 – 2085，2015年。[24] Emmanuel Candes、Yingying Fan、Lucas Janson 和 Jinchi Lv。淘金：用于高维受控变量选择的“model-x”Knockoff。*Journal of the Royal Sta-tistical Society Series B: Statistical Methodology*，80(3):551–577，2018年。[25] Zhimei Ren 和 Emmanuel Candes。利用侧信息的Knockoff。*The Annals ofApplied Statistics*，17(2):1152–1174，2023。[26] Panagiota I Kontou 和 Pantelis G Bagos。GWAS汇总统计量的宝库：方法与工具的系统综述。*BioData Mining*，17(1):31，2024。[27] Yuxin Zou、Peter Carbonetto、Gao Wang 和 Matthew Stephens。利用“单效应之和”模型，根据汇总数据进行精细定位。*PLoS genetics*，18(7):e1010299，2022。

[28] Hui Zou。自适应 Lasso 及其预言机性质。*Journal of the American StatisticalAssociation*，101(476):1418–1429，2006。[29] Zihuai He、Linxi Liu、Michael E Belloy、Yann Le Guen、Aaron Sossin、Xiaoxia Liu、Xinran Qi、Shiyang Ma、Prashnna K Gyawali、Tony Wyss-Coray 等。Ghostknockoff推断助力在全基因组关联研究中识别推定因果变异。*Nature communications*，13(1):7209，2022。[30] Zhaomeng Chen、Zihuai He、Benjamin B Chu、Jiaqi Gu、Tim Morrison、Chiara Sabatti 和 Emmanuel Cand`es。仅凭汇总统计量进行受控变量选择？通过ghostknockoffs和惩罚回归实现。*ArXiv*，第arXiv–2402页，2024。[31] Xiang Zhu 和 Matthew Stephens。基于全基因组关联研究汇总统计量的贝叶斯大规模多重回归。*The annals of applied statistics*，11(3):1561，2017年。[32] Jaime Roquero Gimenez 和 James Zou。提高Knockoff程序的稳定性：多个同步knockoff与熵最大化。载于*The 22nd in-ternational conference on artificial intelligence and statistics*，第2184–2192页。PMLR，2019年。[33] Junyang Qian、Yosuke Tanigawa、Wenfei Du、Matthew Aguirre、Chris Chang、Robert Tibshirani、Manuel A Rivas 和 Trevor Hastie。一个快速且可扩展的框架，用于大规模和超高维稀疏回归，并应用于UK Biobank。*PLoS genetics*，16(10):e1009141，2020年。[34] Qianqian Zhang、Florian Priv´e、Bjarni Vilhj´almsson 和 Doug Speed。利用个体水平数据或汇总统计量改进复杂性状的遗传预测。*Naturecommunications*，12(1):4192，2021年。[35] Matteo Sesia、Chiara Sabatti 和 Emmanuel J Cand`es。利用隐马尔可夫模型knockoffs寻找基因。*Biometrika*，106(1):1–18，2019年。[36] GTEx联盟。GTEx联盟的人类组织遗传调控效应图谱。*Science*，369(6509):1318–1330，2020年。https://doi.org/10.1126/science. aaz1776。[37]Hilary K Finucane、Brendan Bulik-Sullivan、Alexander Gusev、Gosia Trynka、Yakir Reshef、Po-Ru Loh、Verneri Anttila、Han Xu、Chongzhi Zang、Kyle Farh 等。利用全基因组关联研究汇总统计量，根据功能注释划分遗传力。*Nat. Genet.*，47(11):1228–1235，2015年。https://doi.org/10.1038/ng.3404。

[38] S. Gazal. S-ldsc参考文件，2024。数据集。

[39] Roby Joehanes、Xiaoling Zhang、Tianxiao Huan、Chen Yao、Sai-xia Ying、Quang Tri Nguyen、Cumhur Yusuf Demirkale、Michael L Feolo、Nataliya R Sharopova、Anne Stur- cke 等。表达数量性状位点的整合全基因组分析有助于解读基因组关联研究。*Genome Biology*，18(1):16，2017年。[40] Vivek Chandra、Sourya Bhattacharyya、Benjamin J Schmiedel、Ariel Madrigal、Cristian Gonzalez-Colin、Stephanie Fotsing、Austin Crinklaw、Gregory Seumois、Pejman Moham- madi、Mitchell Kronenberg 等。启动子相互作用表达数量性状位点富集了功能性遗传变异。*Nature Genetics*，53(1):110–119，2021年。[41] Lu Chen、Bing Ge、Francesco Paolo Casale、Louella Vasquez、Tony Kwan、Diego Garrido- Mart´ın、Stephen Watt、Ying Yan、Kousik Kundu、Simone Ecker 等。人类免疫细胞中表观遗传和转录变异的遗传驱动因素。*Cell*，167(5):1398–1414，2016年。[42]Elise D Flynn、Athena L Tsu、Silva Kasela、Sarah Kim-Hellmuth、Francois Aguet、Kristin G Ardlie、Harmen J Bussemaker、Pejman Mohammadi 和 Tuuli Lappalainen。转录因子对不同个体和组织中 eQTL 活性的调控。*PLoS Ge-netics*，18(1):e1009719，2022年。[43] Heini M Natri、Christina B Del Azodi、Lance Peter、Chase J Taylor、Sagrika Chugh、Robert Kendle、Mei-i Chung、David K Flaherty、Brittany K Matlock、Carla L Calvi 等。人类肺部细胞类型特异性和疾病相关的表达数量性状位点。*Nature Genetics*，56(4):595–604，2024。[44] Oana Palasca、Alberto Santos、Christian Stolte、Jan Gorodkin 和 Lars Juhl Jensen。Tissues 2.0：哺乳动物组织表达的综合网络资源。*Database*，2018:bay003，2018年。[45] Maxim V Kuleshov、Matthew R Jones、Andrew D Rouillard、Nicolas F Fernandez、Qiao- nan Duan、Zichen Wang、Simon Koplev、Sherry L Jenkins、Kathleen M Jagodnik、Alexan- der Lachmann 等。Enrichr：全面的基因集富集分析网络服务器2016年更新。*Nucleic Acids Research*，44(W1):W90–W97，2016年。[46] Zhuorui Xie、Allison Bailey、Maxim V Kuleshov、Daniel JB Clarke、John E Evangelista、Sherry L Jenkins、Alexander Lachmann、Megan L Wojciechowicz、Eryk Kropiwnicki、Kathleen M Jagodnik 等。利用 Enrichr 进行基因集知识发现。*Current proto-cols*，1(3):e90，2021年。[47]Lo¨ıc Yengo、Sailaja Vedantam、Eirini Marouli、Julia Sidorenko、Eric Bartell、Saori Sakaue、Marielisa Graff、Anders U Eliasen、Yunxuan Jiang、Sridharan Raghavan 等。与人类身高相关的常见遗传变异的饱和图谱。《自然》，610(7933):704–712，2022年。

[48] 千人基因组计划联盟。人类遗传变异的全球参考。《自然》，526(7571):68–74，2015年。https://doi.org/10.1038/nature15393。[49] Brielin C Brown、Chun Jimmie Ye、Alkes L Price 和 Noah Zaitlen。基于汇总统计量的跨族群遗传相关性估计。《美国人类遗传学杂志》，99(1):76–88，2016年。[50] Vassily Trubetskoy、Antonio F Pardi˜ñas、Ting Qi、Georgia Panagiotaropoulou、Swapnil Awasthi、Tim B Bigdeli、Julien Bryois、Chia-Yen Chen、Charlotte A Dennison、Lynsey S Hall 等。基因组位点定位揭示了精神分裂症相关基因和突触生物学。《自然》，604(7906):502–508，2022年。[51] Emil Uffelmann、Qin Qin Huang、Nchangwi Syntia Munung、Jantina De Vries、Yukinori Okada、Alicia R Martin、Hilary C Martin、Tuuli Lappalainen 和 Danielle Posthuma。全基因组关联研究。《自然综述：方法导论》，1(1):59，2021年。[52] Zhimei Ren 和 Rina Foygel Barber。去随机化 Knockoffs：利用 e 值控制错误发现率。《英国皇家统计学会杂志：*B*辑（统计方法论）》，86(1):122–154，2024。https://doi.org/10.1093/jrsssb/qkad085。[53] Kai Yuan、Ryan J Longchamps、Antonio F Pardi˜nas、Mingrui Yu、Tzu-Ting Chen、Shu-Chin Lin、Yu Chen、Max Lam、Ruize Liu、Yan Xia 等。跨多样化祖源精细定位，推动发现人类复杂性状和疾病的推定因果变异。*Nature Genetics*，56(9):1841–1850，2024。[54] Nathan LaPierre、Kodi Taraszka、Helen Huang、Rosemary He、Farhad Hormozdiari 和 Eleazar Eskin。通过跨多项研究精细定位识别因果变异。*PLoS Genetics*，17(9):e1009733，2021年。[55] Mingxuan Cai、Zhiwei Wang、Jiashun Xiao、Xianghong Hu、Gang Chen 和 Can Yang。Xmap：利用遗传多样性并校正混杂偏倚的跨群体精细定位。《自然·通讯》，14(1):6870，2023年。