arXiv:2302.01982v1 [stat.ME] 3 Feb 2023

multi-GPA-Tree：一种基于多效性信息和功能注释树引导全基因组关联研究结果优先级排序的统计方法

AasthaKhatiwada[1], AyseSelen Yilmaz[2], Bethany J.Wolf[3], Maciej Pietrzak[2], 以及 Dongjun Chung[2,4,*]

[1]National Jewish Health 生物统计学与生物信息学系，美国科罗拉多州丹佛[2]俄亥俄州立大学生物医学信息学系，美国俄亥俄州哥伦布[3]南卡罗来纳医科大学公共卫生科学系，美国南卡罗来纳州查尔斯顿[4]俄亥俄州立大学詹姆斯综合癌症中心 Pelotonia免疫肿瘤学研究所*通讯作者（chung.911@osu.edu）。

摘要

全基因组关联研究（GWAS）已成功发现了二十多万个基因型-性状关联，但仍面临一些挑战。首先，复杂性状通常与许多单核苷酸多态性（SNP）相关，其中大多数效应大小较小或中等，因此难以检测。其次，许多复杂性状因“多效性”而具有共同的遗传基础。尽管考虑多效性的方法不多，但利用多效性能够提高检测效应较弱的基因型-性状关联的统计功效。第三，目前可用的统计方法在解释遗传变异与特定或多个性状相关联的功能机制方面能力有限。我们提出 multi-GPA-Tree 来应对这些挑战。multi-GPA-Tree 方法能够识别与单一性状及多个性状相关的风险单核苷酸多态性，同时还能识别可解释风险相关单核苷酸多态性与这些性状之间关联机制的功能注释组合。

首先，我们开展了模拟研究，以评估所提出的 multi-GPA-Tree 方法，并将其性能与现有统计方法进行了比较。结果表明，在检测多个性状的风险相关单核苷酸多态性方面，multi-GPA-Tree 优于现有统计方法。其次，我们将 multi-GPA-Tree 应用于系统性红斑狼疮（SLE）与类风湿关节炎（RA）的全基因组关联研究，以及克罗恩病（CD）与溃疡性结肠炎（UC）的全基因组关联研究，并使用了包括 GenoSkyline 和 GenoSkylinePlus 在内的功能注释数据。我们的结果表明，multi-GPA-Tree 可成为一种强有力的工具，既能改进关联定位，也有助于理解复杂性状的遗传结构，以及风险相关单核苷酸多态性与复杂性状之间联系的潜在机制。

可用性：multiGPATree软件可在 https://dongjunchung.github.io/multiGPATree/ 获取。

1 引言

随着人们对识别与复杂性状相关的基因组区域日益关注，已发表的全基因组关联研究（GWAS）及基因型-性状关联数量大幅增加（https://www.ebi.ac.uk/gwas/）[1]。此前未知的基因型-性状关联得到识别后，许多复杂性状的遗传力（性状内的遗传变异）估计也得到了改善。然而，仍然存在两项主要挑战。首先，由于尚未识别出的基因型-性状

关联 [2, 3, 4]，许多性状仍有一部分遗传力尚未得到解释。多基因性是一种遗传变异与性状之间仅存在较弱或中等效应大小关联的现象 [5, 6] ，可合理解释遗传力缺失。理论上，可以通过扩大GWAS样本量，提高检测弱关联和中等关联的统计功效，从而减轻多基因性的影响；然而，招募大量样本往往需要更多资源，而且由于性状在人群中的患病率有限，并非总是可行 [7]。除扩大样本量以提高统计功效外，还可以同时整合多个性状的全基因组关联研究关联汇总统计量，利用两个或多个性状之间的多效性关系（共享遗传基础） [8, 9, 10]。GWAS汇总统计可通过公共数据存储库轻松获取并用作输入 [1, 11] ，可作为个体水平基因型-表型数据的良好替代；后者则更难获取。其次，大多数GWAS识别出的位点都位于非编码区域 [12], ，因此难以理解与已识别的基因型-性状关联相关的功能机制。例如，在自身免疫性疾病中，约90%的因果遗传变异位于非编码区域，其中大部分位于调控性DNA区域 [14, 13]。因此，利用基因组功能注释信息（可提供不同类型的组蛋白修饰、表观遗传变化以及细胞和组织特异性变化等相关信息），有助于解读将风险相关遗传变异与性状联系起来的功能机制 [16, 17, 15]。因此，将遗传数据与基因组功能注释数据进行整合分析是一个很有前景的方向。

建立在数据整合方法基础上的统计方法，不仅利用了公共数据存储库中现成的信息，还克服了多基因性带来的挑战，同时能够揭示与一个或多个性状相关的潜在功能机制。因此，这些方法更具优势，也更为高效。认识到数据整合方法有望提高关联检测的统计功效后，研究人员开发了若干针对GWAS汇总统计的统计方法 [16, 17, 8, 10, 19, 20]。这些方法大体可分为三类。

第一类方法通过同时整合多个GWAS关联 p值，利用两个或多个不同性状之间的多效性关系。该类别中两种备受青睐的方法是多效性信息条件错误发现率方法 [8] 和图GPA方法[10]。这两种方法的共同目标是提高统计功效，从而优先筛选出与一个或多个性状风险相关的单核苷酸多态性。条件错误发现率方法提高了对两种精神障碍——精神分裂症和双相情感障碍——的风险相关单核苷酸多态性的检出能力。尽管该方法易于实施，但在估算条件错误发现率时缺乏基于模型的框架，这会削弱检测非零关联以及推断非零分布特征的能力。此外，该方法只能整合少量GWAS性状。相比之下，graph-GPA采用隐马尔可夫随机场框架，可整合大量GWAS性状；其有效性已通过整合12种性状（五种精神障碍、三种自身免疫性状、两种脂质相关性状和两种心血管性状）得到证明，其中临床相关性状形成了紧密相连的簇。然而，由于无法在应用中整合功能注释，这两种方法都无法揭示风险相关单核苷酸多态性的功能相关性。

第二类方法将个体全基因组关联研究数据与基因型相关功能注释数据相结合。该类方法中两种前沿方法包括潜在稀疏混合模型（LSMM）方法 [16] 和协变量调节的假发现率（cmFDR）方法 [17]。在LSMM中，采用逻辑混合效应模型框架整合功能注释，其中假定基因和细胞类型特异性功能注释分别具有固定效应和随机效应，并对随机效应施加稀疏结构，以自适应地筛选可能与性状病因学相关的细胞类型特异性功能注释。通过应用LSMM，Ming等人发现，系统性红斑狼疮、类风湿关节炎、溃疡性结肠炎和克罗恩病等自身免疫性疾病显著富集了血液相关的细胞类型特异性注释。类似

于LSMM，cmFDR方法是一种参数方法，它整合GWAS汇总统计和功能注释信息，其中功能注释信息在参数双组混合模型中提供“先验信息”。cmFDR方法假设，与功能上无关的单核苷酸多态性相比，功能上相关的单核苷酸多态性具有较低的假发现率，且与该性状相关。然而，cmFDR和LSMM均未利用病因学相似的性状之间的多效性关系来提高关联检测能力。

最后，第三类统计方法结合了前两类方法的标准，将多个 GWAS 性状数据与基因型相关功能注释数据整合起来。该类别中两种广为人知的方法包括整合多效性与注释的遗传分析（GPA）方法 [19] 以及较新的潜变量概率单位模型（LPM）方法 [20]。GPA 利用多效性和功能注释数据，对多个性状进行联合分析，从而采用统一统计框架整合具有遗传相关性的 GWAS 性状。与 GPA 类似，LPM 的三个主要目标是：通过估计多个性状之间的相关性来识别其多效性关系；识别功能注释的影响；以及提高识别一个或多个性状风险相关单核苷酸多态性的能力。在这两种方法中，随着 GWAS 性状和功能注释数量增加，模型中包含的参数数量也会显著增加，使得其在统计和计算层面的实现面临挑战。此外，尽管第二类和第三类方法可以对单项注释进行富集分析，但这些方法并未考虑注释之间的交互作用，因此在阐明遗传变异与一个或多个性状相关联所涉及的组合功能通路方面存在局限。虽然其中一些方法理论上可以扩展为纳入功能注释之间的交互作用，以评估注释的综合功能效应，但仍需预先了解所关注的交互作用。因此，亟需一种能够进行变量选择的方法，从大量与遗传变异相关的注释中，识别出与一个或多个性状相关的功能注释或其组合。

为应对上述统计学挑战与局限性，我们团队近期发表了一种名为 GPA-Tree 的新型统计方法 [21] ，可同时进行关联定位并识别功能注释之间的交互作用。然而，GPA-Tree 未利用两个或更多性状之间的多效性关系来提高关联定位的统计功效。在本研究中，我们提出了一种名为 multi-GPA-Tree 的新方法，以解决 GPA-Tree方法的局限性。multi-GPA-Tree 方法是一种基于分层建模架构、结合多变量回归树算法 [22] 的新型统计方法。该方法利用具有相似病因学的性状之间的多效性关系，在优先确定一个或多个性状相关单核苷酸多态性的同时，识别与相关机制有关的关键功能注释组合，即这些机制决定了一个或多个性状相关单核苷酸多态性如何影响性状。全面的模拟研究和真实数据应用表明，multi-GPA-Tree 能够持续提高检测一个或多个性状相关单核苷酸多态性的统计功效，并有效识别具有重要生物学意义的功能注释组合。multi-GPA-Tree 方法以多个性状的 GWAS汇总统计和 GWAS遗传变异的功能注释信息作为输入，并可使用 R软件包“multiGPATree”实现。

材料与方法

multi-GPA-Tree方法概述

设$\mathbf{Y}_{M\times D}$为一个基因型-性状关联 p值矩阵，其中包含 $i = 1, 2、\cdots、 M$个单核苷酸多态性和 d =1、2、$\cdots、 D$个性状，其中 Y_{id}表示 i^{th}单核苷酸多态性与 d^{th}性状关联的 p值。

$$\mathbf{Y} = (\mathbf{Y}_{.1}, \ldots, \mathbf{Y}_{.D}) = \begin{pmatrix} y_{11} & \cdots & y_{1D} \\ \vdots & \ddots & \vdots \\ y_{M1} & \cdots & y_{MD} \end{pmatrix}_{M\times D}$$

我们还假设每个单核苷酸多态性（SNP）有 K 个二元注释（$\mathbf{A}$）。

$$\mathbf{A} = (\mathbf{A}_{.1}, \ldots, \mathbf{A}_{.K}) = \begin{pmatrix} a_{11} & \cdots & a_{1K} \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MK} \end{pmatrix}_{M\times K}, \quad \text{where}$$

$$a_{ik} = \begin{cases} 0, & \text{if } i^{th} \text{ SNP is not annotated in the } k^{th} \text{ annotation} \\ 1, & \text{if } i^{th} \text{ SNP is annotated in the } k^{th} \text{ annotation} \end{cases}$$

为了提高识别与一种或多种性状相关的风险相关单核苷酸多态性（SNP）的能力，我们将针对 D 种性状（$\mathbf{Y}$）的全基因组关联研究（GWAS）关联 p- 值与功能注释数据（$\mathbf{A}$）整合起来。功能注释在模拟全基因组关联研究（GWAS）性状与单核苷酸多态性（SNP）之间关系时的影响，通过定义由潜在二元变量构成的矩阵 $\mathbf{Z}_{M\times 2^D} \in \{0,1\}$ 来表征，其中 $\mathbf{Z}_i$ 是一个长度为 2^D 的向量，表示该 i^{th} 单核苷酸多态性（SNP）对于这些 D 性状是零效应还是非零效应。为简化符号表示，我们在此介绍两个全基因组关联研究（GWAS）性状（$D = 2$）情况下的模型。

设 $Y \in \mathbf{R}^{M\times 2}$ 为两个性状的全基因组关联研究（GWAS）关联 p-值矩阵，其中 Y_{i1} 和 Y_{i2} 分别是该 i^{th} 单核苷酸多态性（ SNP ）与性状 1 和性状 2 关联的 p- 值。对于该 i^{th} 单核苷酸多态性（ SNP ），潜在二元向量定义为 $\mathbf{Z}_i = \{Z_{i00}, Z_{i10}, Z_{i01}, Z_{i11}\}$，其中 $Z_{i00} = 1$ 表示该 i^{th} 单核苷酸多态性（SNP）对两个性状均为零效应， $Z_{i10} = 1$ 表示该 i^{th} 单核苷酸多态性（SNP）对性状 1 为非零效应、对性状 2 为零效应， $Z_{i01} = 1$ 表示该 i^{th} 单核苷酸多态性（SNP）对性状 1 为零效应、对性状 2 为非零效应，而 $Z_{i11} = 1$ 表示该 i^{th} 单核苷酸多态性（SNP）对两个性状均为非零效应。我们假设一个单核苷酸多态性（SNP）只能处于四种状态之一，即 $\sum_{l\in\{00,10,01,11\}} Z_{il} = 1$。

假设两种性状中，零组和非零组的单核苷酸多态性密度分别服从 $U[0,1]$和 $Beta$（α_d，1）分布，其中 $0 < \alpha_d < 1$和 $d = 1, 2$，正如 Chung 等人[19]所提出的。分布定义如下。

$$\begin{aligned} &Y_{i1}|Z_{i00} = 1 \sim U[0,1] && Y_{i2}|Z_{i00} = 0 \sim U[0,1] \\ &Y_{i1}|Z_{i10} = 1 \sim Beta(\alpha_1, 1) && Y_{i2}|Z_{i10} = 1 \sim U[0,1] \\ &Y_{i1}|Z_{i01} = 1 \sim U[0,1] && Y_{i2}|Z_{i01} = 1 \sim Beta(\alpha_2, 1) \\ &Y_{i1}|Z_{i11} = 1 \sim Beta(\alpha_1, 1) && Y_{i2}|Z_{i11} = 1 \sim Beta(\alpha_2, 1), \end{aligned}$$

其中 $0 < \alpha_1, \alpha_2 < 1$。最后，通过定义函数 f，将功能注释数据 $\mathbf{A}$ 与全基因组关联研究汇总统计数据 $\mathbf{Y}$ 整合起来；该函数由功能注释$\mathbf{A}$ 构成，并将其与式 1 中给出的潜在 $\mathbf{Z}$ 的多变量期望相关联。

$$P(Z_{il} = 1; a_{i1}, \ldots, a_{iK}) = f(a_{i1}, \ldots, a_{iK}), \text{where } l \in \{00, 10, 01, 11\} \tag{1}$$

为方便记号，我们令 $\boldsymbol{\theta} = (\alpha_1, \alpha_2)$，并将 $P(Z_{il} = 1; a_{i1}, \ldots\ldots, a_{iK})$ 记为 $\boldsymbol{\pi}_{.l}$，其中

$l \in \{00, 10, 01, 11\}$ 其中， $\boldsymbol{\pi}_{.00}$ 表示单核苷酸多态性对两种性状均为零效应的先验概率， $\boldsymbol{\pi}_{.10}$表示单核苷酸多态性对性状 1 为非零效应、对性状 2 为零效应的先验概率， $\boldsymbol{\pi}_{.01}$ 表示单核苷酸多态性对性状 1 为零效应、对性状 2 为非零效应的先验概率， $\boldsymbol{\pi}_{.11}$ 表示单核苷酸多态性对两种性状均为非零效应的先验概率。随后，假设单核苷酸多态性彼此独立，则观测数据 $Pr(\mathbf{Y}, \mathbf{A})$ 的联合分布，以及不完整数据和完整数据对数似然，可分别写成公式 2、3 和 4 所示的形式。

$$\begin{aligned} Pr(\mathbf{Y}, \mathbf{A}) &= \prod_{i=1}^{M} \left[\sum_{l \in \{00,10,01,11\}} P(Z_{il} = 1) P(Y_{i1}, Y_{i2} | Z_{il} = 1) \right] \\ &= \prod_{i=1}^{M} \left[\sum_{l \in \{00,10,01,11\}} \pi_{il} \, P(Y_{i1}, Y_{i2} | Z_{il} = 1) \right] \end{aligned} \tag{2}$$

$$\ell_{IC} = \sum_{i=1}^{M} log \left[\sum_{l \in \{00,10,01,11\}} \pi_{il} \, P(Y_{i1}, Y_{i2} | Z_{il} = 1) \right] \tag{3}$$

$$\ell_{C} = \sum_{i=1}^{M} \sum_{l \in \{00,10,01,11\}} Z_{il} \, log \left[\pi_{il} \, P(Y_{i1}, Y_{i2} | Z_{il} = 1) \right] \tag{4}$$

算法

根据上述方法，参数估计采用期望最大化（EM）算法 [23]实现。通过多变量回归树算法[22] 估计式 1 中的函数 f ，该算法能够识别与特定性状和多个性状的风险相关单核苷酸多态性相关的功能注释组合。根据模拟研究的发现，采用两阶段方法可改善参数估计和模型稳定性，因此所述方法在计算上分为两个阶段实现。具体而言，在阶段1，我们首先估计参数 α_1 和 α_2 ，而不识别功能注释组合。随后，在阶段2，我们识别关键的功能注释组合（f(**A**)），同时将参数 α_1 和 α_2 固定为阶段1中获得的值。下文将详细说明计算步骤。

阶段1： 在阶段1中，我们初始化 $\alpha_d^{(0)} = 0.1$、 $d = 1,2$ 和 $\pi_{il}^{(0)} = \frac{1}{2^D}$、 $D = 2$ （性状数量）。在第t^{th}次迭代的 E步中，为该单核苷酸多态性定义 $Z_{il}^{(t)}$, $l \in \{00, 10, 01, 11\}$：

$$\begin{aligned} \mathbf{E-step}: z_{il}^{(t)} &= P(Z_{il} = 1 | \mathbf{Y}, \mathbf{A}; \, \boldsymbol{\theta}^{(t-1)}) \\ &= \frac{\pi_{il}^{(t-1)} \, P(Y_{i1}, Y_{i2} | Z_{il} = 1; \boldsymbol{\theta}^{(t-1)})}{\sum\limits_{l' \in \{00,10,01,11\}} \pi_{il'}^{(t-1)} \, P(Y_{i1}, Y_{i2} | Z_{il'} = 1; \boldsymbol{\theta}^{(t-1)})} \end{aligned} \tag{5}$$

在 M步的第 t^{th} 次迭代中， $\boldsymbol{\pi}_{i.}$、 α_1 和 α_2 更新如下：

M –步： 拟合多变量线性回归模型如下

$$\mathbf{Z}_{i.}^{(t)} = \beta_0^{(t)} + \beta_1^{(t)} a_{i1} + \cdots + \beta_K^{(t)} a_{iK} + \epsilon_i{}^{(t)}$$

将 $\boldsymbol{\pi}_{i.}$ 更新为多变量线性回归模型的预测值 .

$$\text{Update } \alpha_1^{(t)} = -\frac{\sum\limits_{i=1}^{M} (z_{i10}^{(t)} + z_{i11}^{(t)})}{\sum\limits_{i=1}^{M} (z_{i10}^{(t)} + z_{i11}^{(t)})(logY_{i1})} \text{ and } \alpha_2^{(t)} = -\frac{\sum\limits_{i=1}^{M} (z_{i01}^{(t)} + z_{i11}^{(t)})}{\sum\limits_{i=1}^{M} (z_{i01}^{(t)} + z_{i11}^{(t)})(logY_{i2})}$$

其中 $\beta_k^{(t)}$、 $k = 0$、 $\cdots$、 K 为回归系数， $\epsilon_i^{(t)}$ 为误差项。重复 E 步和 M步，直到不完全对数似然以及 α_1 和 α_2 估计值收敛。然后，使用本阶段估计的 α_1、 α_2 和 $\boldsymbol{\pi}_{i.}$ 分别固定阶段2中的 α_1、 α_2 并初始化 $\boldsymbol{\pi}_{i.}$。

阶段2： I 在阶段2中，我们采用多变量回归树算法实现另一个EM算法 e

算法 ，从而能够识别注释之间的并集、交集和补集关系 n

功能注释来估计 $\boldsymbol{\pi}_{i.}$。在E步的第 t^{th} 次迭代中，如公式5所示，为 i^{th} 单核苷酸多态性定义 $Z_{il}^{(t)}$，$l \in \{00, 10, 01, 11\}$，但 α_1 和 α_2 固定为$\hat{}\alpha_1$ 和$\hat{}\alpha_2$，它们分别是从阶段1获得的 α_1 和 α_2 的最终估计值。

$$\mathbf{E-step}: \text{Define } Z_{il}^{(t)},\ l \in \{00, 10, 01, 11\} \text{ as in Eq 5, except } \alpha_1 \text{ and } \alpha_2 \text{ are fixed as } \hat{\alpha_1} \text{ and } \hat{\alpha_2}, \text{ the final estimates of } \alpha_1 \text{ and } \alpha_2 \text{ from Stage 1.}$$

在 t^{th} M步迭代中， $\boldsymbol{\pi}_{i.}$ 更新如下：

$$\mathbf{M-步}: \text{拟合如下所示的多变量回归树模型。} \mathbf{Z}_{i.}^{(t)} = f^{(t)}(a_{i1}, \cdots, a_{iK}) + \epsilon_i^{(t)}, \text{其中 } \epsilon_i \text{ 为误差项。将 } \boldsymbol{\pi}_{i.}{}^{(t)} \text{ 更新为多变量回归树模型的预测值。} \tag{6}$$

在M步中，多变量回归树的复杂度参数（cp）是关键调优参数，其定义为树的每个节点所需的最小改进幅度。具体而言，在多变量回归树模型中，首先构建可能的最大规模树（即完整规模树），然后使用 cp对其进行剪枝。该方法能够构建既准确又具有可解释性的多变量回归树，用于解释功能注释与一个或多个性状的风险相关单核苷酸多态性之间的关系。E步和M步重复进行，直到不完全对数似然收敛。多变量回归树模型在阶段2 EM收敛时确定的剪枝树结构，即为公式1中的 f。

我们注意到，与标准EM算法不同，阶段2的不完全对数似然不保证单调递增。因此，我们将阶段2实现为广义EM算法，仅保留不完全对数似然高于前一次迭代的迭代结果。

边际和联合风险相关SNP的优先排序及相关功能注释的识别

参数估计完成后，我们可以使用局部错误发现率或 fdr对一个或多个性状的风险相关SNP进行优先排序。如公式7所示，对于特定性状的边际关联，我们将 fdr定义为：在给定该SNP针对所有性状的全基因组关联研究关联 p值及功能注释信息的条件下，该 i^{th} SNP属于该特定性状非风险相关组的边际后验概率。同样，对于性状间的联合关联，我们将 fdr 定义为：在给定该SNP针对所有性状的全基因组关联研究关联 p值及功能注释信息的条件下，该 i^{th}SNP属于这些性状非风险相关组的联合后验概率。接下来，我们采用“直接后验概率”方法[24] 来控制全局错误发现率（FDR）。

$$\begin{aligned}
&fdr_1(\mathbf{Y}_{i.}, \mathbf{A}_{i.}) = P(Z_{i00} + Z_{i01} = 1 | \mathbf{Y}_{i.}, \mathbf{A}_{i.}, \hat{\boldsymbol{\theta}}) = \frac{P(Y_{i1}, Y_{i2}, Z_{i00} + Z_{i01} = 1; \hat{\boldsymbol{\theta}})}{P(Y_{i1}, Y_{i2}; \hat{\boldsymbol{\theta}})},\\
&fdr_2(\mathbf{Y}_{i.}, \mathbf{A}_{i.}) = P(Z_{i00} + Z_{i10} = 1 | \mathbf{Y}_{i.}, \mathbf{A}_{i.}, \hat{\boldsymbol{\theta}}) = \frac{P(Y_{i1}, Y_{i2}, Z_{i00} + Z_{i10} = 1; \hat{\boldsymbol{\theta}})}{P(Y_{i1}, Y_{i2}; \hat{\boldsymbol{\theta}})},\\
&fdr_{1,2}(\mathbf{Y}_{i.}, \mathbf{A}_{i.}) = P(Z_{i00} + Z_{i10} + Z_{i01} = 1 | \mathbf{Y}_{i.}, \mathbf{A}_{i.}) = \frac{P(Y_{i1}, Y_{i2}, Z_{i00} + Z_{i10} + Z_{i01} = 1; \hat{\boldsymbol{\theta}})}{P(Y_{i1}, Y_{i2}; \hat{\boldsymbol{\theta}})},\\
&\text{where}\\
&P(Y_{i1}, Y_{i2};\ \hat{\boldsymbol{\theta}}) = \sum_{l \in \{00,10,01,11\}} \hat{\pi}_{il}\ P(Y_{i1}, Y_{i2} | Z_{il}, \mathbf{A}_{i.};\ \hat{\boldsymbol{\theta}}),\\
&P(Y_{i1}, Y_{i2}, Z_{i00} + Z_{i01} = 1;\ \boldsymbol{\theta}) = \sum \hat{\pi}_{il}\ P(Y_{i1}, Y_{i2} | Z_{il}, \mathbf{A}_{i.};\ \boldsymbol{\theta}), l \in \{00,01\}\hat{}\ \hat{}P(\\
&Y_{i1}, Y_{i2}, Z_{i00} + Z_{i10} = 1;\ \boldsymbol{\theta}) = \sum \hat{\pi}_{il}\ P(Y_{i1}, Y_{i2} | Z_{il}, \mathbf{A}_{i.};\ \boldsymbol{\theta}), l \in \{00,10\}\hat{}\ \hat{}P(Y_{i1},\\
&Y_{i2}, Z_{i00} + Z_{i10} + Z_{i01} = 1;\ \boldsymbol{\theta}) = \sum \hat{\pi}_{il}\ P(Y_{i1}, Y_{i2} | Z_{il}, \mathbf{A}_{i.};\ \boldsymbol{\theta}), l \in \{00,10,01\}
\end{aligned} \tag{7}$$

最后，在第二阶段 EM 算法收敛后，基于多变量回归树模型所选定的功能注释组合，推断出相关的功能注释组合。

结果

模拟研究

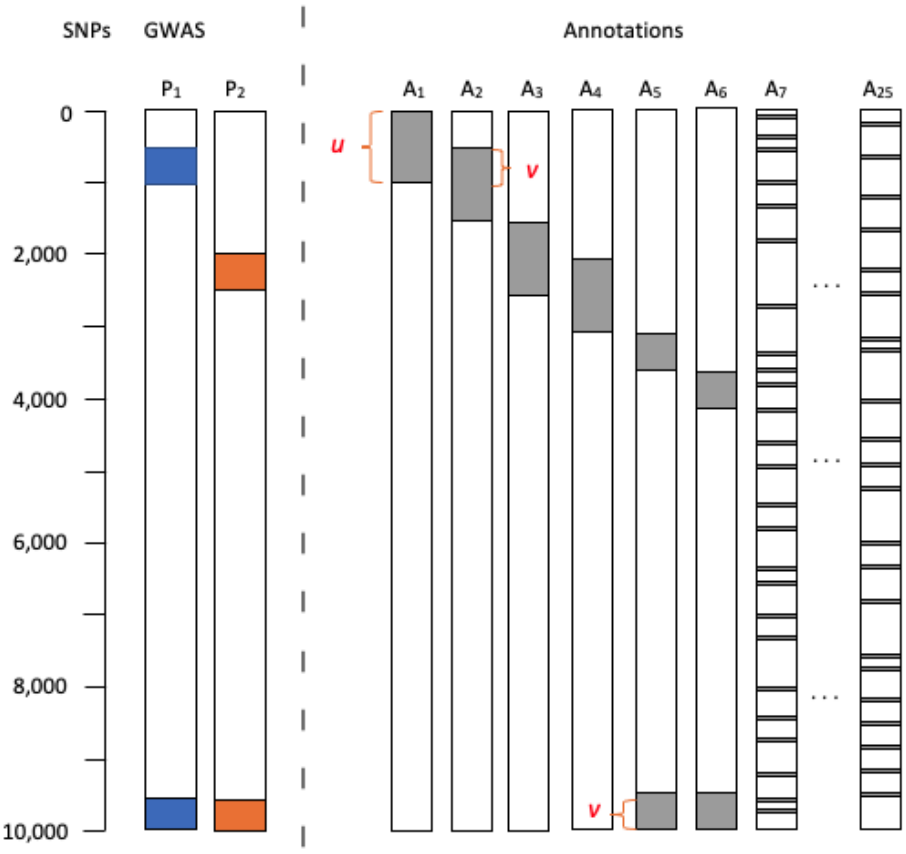

图 1：**模拟设置**。 图示场景展示了 $M = 10,000$ 个单核苷酸多态性； $K = 25$ 个注释； A_1-A_6 (u) 中经注释的单核苷酸多态性占 % $= 10\%$； A_1-A_2、 A_3-A_4、 A_5-A_6 之间的重叠占 % (v)=50%； A_7–A_{15} 为噪声单核苷酸多态性，其中约 20% 被随机注释；蓝色单核苷酸多态性对性状 P_1 非零，其全基因组关联研究 p 值由 $Beta(\alpha_1 = 0.4, 1)$ 分布生成；橙色单核苷酸多态性对性状 P_2 非零，其全基因组关联研究 p 值由 $Beta(\alpha_2 = 0.4, 1)$ 分布生成；所有其他单核苷酸多态性对两种性状均为零，其全基因组关联研究 p 值对两种性状均由 $U[0, 1]$ 分布生成。

我们开展了一项模拟研究，以评估所提出的 multi-GPA-Tree 方法的性能。图 1 展示了模拟设置的示意图。对于所有模拟数据，单核苷酸多态性的数量设为 $M = 10{,}000$，注释的数量设为 $K = 25$；假定与第一种性状 (P_1) 存在边际关联的单核苷酸多态性具有由 $L_1 = A_1 \cap A_2$ 定义的功能注释组合，假定与第二种性状 (P_2) 存在边际关联的单核苷酸多态性具有由 $L_2 = A_3 \cap A_4$ 定义的功能注释组合，假定与性状 P_1 和 P_2 存在联合关联的单核苷酸多态性具有由 $L_3 = A_5 \cap A_6$ 定义的功能注释组合；其余所有功能

注释 ($A_k, k = 7,...,25$) 均视为噪声注释。假定约 10% 的单核苷酸多态性被注释为注释 $A_1 - A_6$，而 v% 中的 $v = 35$%,50% 和 75% 的已注释单核苷酸多态性在真实功能注释组合之间存在重叠。对于噪声注释 $A_7 - A_{25}$，先从 $Unif[0.1, 0.3]$ 生成已注释单核苷酸多态性的比例，再随机将该比例的单核苷酸多态性设为 1；假定性状 P_1 中满足 L_1 或 L_3 所示功能注释组合的单核苷酸多态性为风险相关单核苷酸多态性，其 p 值从 $Beta(\alpha_1, 1)$ 中模拟得到，参数为 $\alpha_1 = 0.4$。类似地，假定性状 P_2 中满足 L_2 或 L_3 所示功能注释组合的单核苷酸多态性为风险相关单核苷酸多态性，其 p 值从 $Beta(\alpha_2, 1)$ 中模拟得到，参数为 $\alpha_2 = 0.3$。假定不满足与 P_1 或 P_2 相关的必要条件的单核苷酸多态性为非风险单核苷酸多态性，其 p 值从 $U[0, 1]$ 中模拟得到。

我们模拟了 50 个数据集，并使用上述定义的模拟参数比较 multi-GPA-Tree 与 LPM [20] 的性能。用于比较这些方法的指标包括：（1）边际关联和联合关联的曲线下面积（AUC），该曲线通过绘制真正率（灵敏度）与假正率（1-特异度）得到，用于在不同全局错误发现率水平下检测一个或多个性状的风险相关单核苷酸多态性；（2）在局部假发现率（$lfdr$）控制在名义水平 0.20 时，识别边际和联合风险相关单核苷酸多态性的统计功效；（3）当 $lfdr$ 控制在名义水平 0.20 时，预测的 $lfdr$；以及（4）用于生成性状 P_1 和 P_2风险相关组 p-值的 $Beta(\alpha_d, 1)$ $d = 1, 2$ 分布中 α_d 参数的估计准确度。对于 multi-GPA-Tree，我们还考察了正确检测功能注释树的准确性，具体指标包括：（1）所有相关功能注释在 L_1、L_2 和 L_3（即注释 $A_1 - A_6$）中均被同时识别的数据集所占比例；（2）multi-GPA-Tree 识别出的功能注释中，噪声功能注释（$A_7 - A_{25}$）的平均比例；以及（3）multi-GPA-Tree 识别出的功能注释中，真实功能注释（$A_1 - A_6$）的平均比例。特别地，我们考察了在 $A_1 - A_2$、$A_3 - A_4$ 和 $A_5 - A_6$ 中注释的单核苷酸多态性之间的重叠（v）如何影响 multi-GPA-Tree 将相关功能注释与噪声注释区分开来、以检测一个或多个性状风险相关单核苷酸多态性的能力。

- **曲线下面积：** 图2A比较了 multi-GPA-Tree 和 LPM 得出的曲线下面积分布。对于所有v，在边际关联和联合关联方面，multi-GPA-Tree 的曲线下面积均始终高于 LPM。LPM 在边际关联方面的曲线下面积高于联合关联。

- **统计功效：** 图2B比较了在局部错误发现率（$lfdr$）控制在 0.20 时，multi-GPA-Tree 和 LPM 检测真实边际和联合风险相关单核苷酸多态性的功效分布。对于所有 v，与 LPM 相比，multi-GPA-Tree 方法检测真实边际和联合风险相关单核苷酸多态性的统计功效更高。LPM 在边际关联方面的功效高于联合关联。总体而言，LPM 的统计功效变异性高于 multi-GPA-Tree，而在较高的 v下，multi-GPA-Tree 的功效变异性更大。

- **预测局部错误发现率（lfdr）**：图2C比较了在 $lfdr$ 控制于名义水平 0.20 时，multi-GPA-Tree 和 LPM 的预测 $lfdr$ 分布。尽管 LPM 的预测 $lfdr$ 高于 multi-GPA-Tree，但对于所有 v，在 0.20 水平下，multi-GPA-Tree 和 LPM 的 $lfdr$ 均始终控制在 0.20 以下。

- **参数估计 α：** 图2D展示了使用 multi-GPA-Tree 和 LPM 对性状 1 和 2（P1 和 P2）的 α 参数估计值分布。对于这两种性状，LPM 的平均估计准确度高于 multi-GPA-Tree。multi-GPA-Tree 方法通常会高估 α，在 v较小时尤为明显。随着 v 增加，multi-GPA-Tree 得出的 α 估计值逐渐接近真实值。值得注意的是，multi-GPA-Tree 对 α 的高估并未影响其

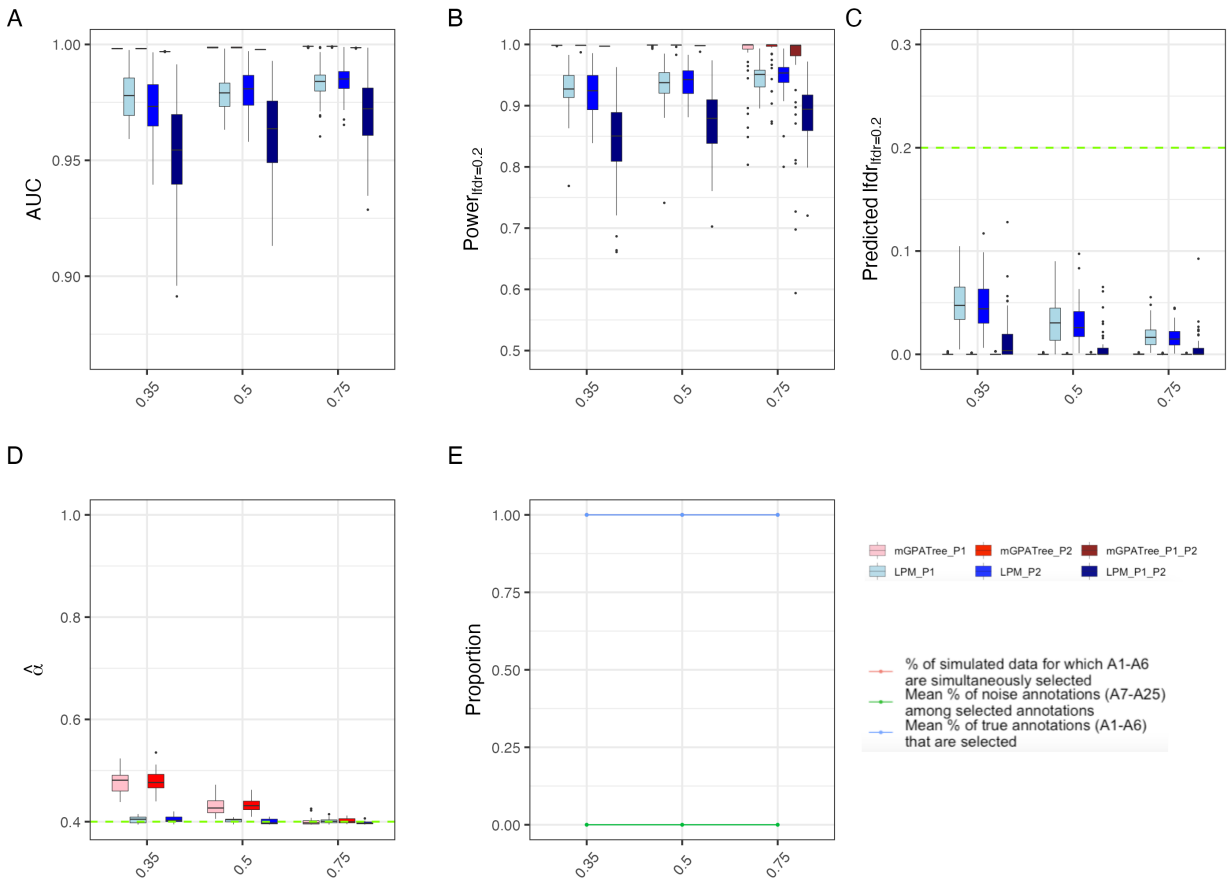

图2：**模拟结果**。比较 multi-GPA-Tree 和 LPM 在性状 P1 和 P2 上的 (A) 曲线下面积，(B) 当局部假发现率（$lfdr$）控制在名义水平 0.20 时检测真实边际关联和联合风险相关单核苷酸多态性的统计功效，(C) 控制在名义水平 0.20 时预测的局部假发现率（$lfdr$），以及 (D) 估计的 α_1 和 α_2 参数；(E) 仅真实注释（$A_1 - A_6$）被同时选中的模拟数据比例（红线）、multi-GPA-Tree 识别出的功能注释中噪声注释（$A_7 - A_{25}$）的平均比例（绿线），以及 multi-GPA-Tree 识别出的注释中真实注释 $A_1 - A_6$ 的平均比例（蓝线）。注意，图中蓝线与红线重叠。结果按在 $A_1 - A_2$、$A_3 - A_4$ 和 $A_5 - A_6$ 中注释的单核苷酸多态性重叠比例（v；x 轴）分别展示。 $M = 10,000$、$K = 25$以及$\alpha_d = 0.4$在 $Beta(\alpha_d, 1)$ $d = 1, 2$。结果汇总自 50 次重复实验。边际关联相关结果使用后缀 * P1 和 * P2 表示，联合关联相关结果使用后缀 * P1 P2 表示。

识别真实功能注释组合或边际及联合风险相关单核苷酸多态性的能力，而这些正是 multi-GPA-Tree 的主要目标。

- **相关与噪声注释的选择：**图 2E 中的红线表示 multi-GPA-Tree 同时识别出仅由真实组合 L_1, L_2 和 L_3 ($A_1 - A_6$) 中的功能注释构成的比例。红线与蓝线完全重合；蓝线表示所有选定注释中真实注释的平均比例（$A_1 - A_6$）。最后，绿线表示选定注释中噪声注释的比例（$A_7 - A_{25}$）。红线和蓝线重合，且绿线保持在 0 的水平，说明 multi-GPA-Tree 选出了且仅选出了相关注释。这些结果展示了.....的潜力

multi-GPA-Tree 能够正确识别真实注释与噪声注释。

真实数据应用

我们获得了一个合并数据集，其中包含系统性红斑狼疮 [30] 、类风湿关节炎 [31], 、克罗恩病和溃疡性结肠炎 [32] 的全基因组关联研究。系统性红斑狼疮和类风湿关节炎全基因组关联研究的汇总统计量分别来自具有欧洲血统的 18,264 名（6, 748 名病例和 11, 516 名对照）和 58,284 名（14, 361 名病例和 43, 923 名对照）个体。克罗恩病和溃疡性结肠炎全基因组关联研究的汇总统计量来自具有欧洲血统的 8, 467 名个体（4, 686 名病例和 3,781 名对照）。经过质量控制并排除主要组织相容性复合体区域内的单核苷酸多态性后，最终分析使用了约 492,557 个单核苷酸多态性，并整合了来自 GenoSkyline（GS） [33] 和 GenoSkylinePlus（GSP） [34] 的功能注释数据。四项全基因组关联研究数据的曼哈顿图和 p 值直方图分别见图 3A 和 3B。

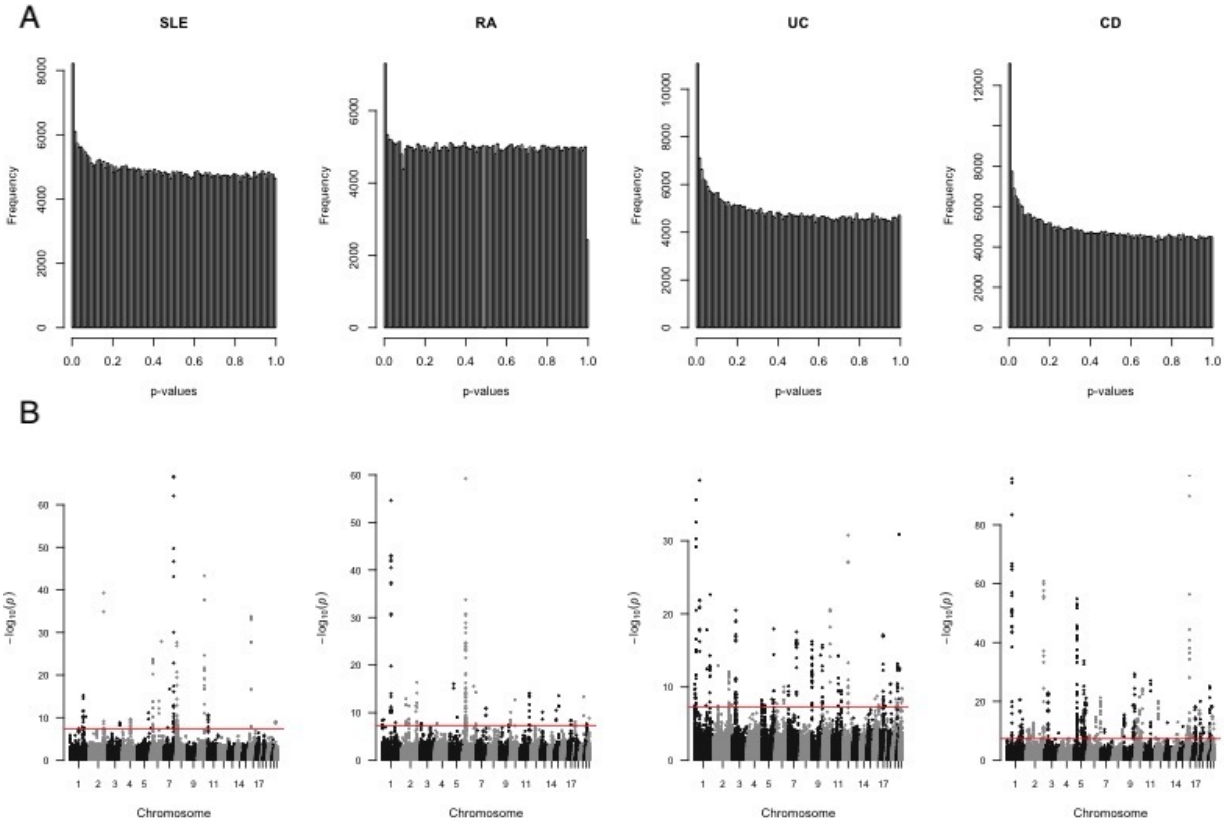

图 3：**GWAS 汇总统计量图**。(A) 四项 GWAS 的 p 值直方图和 (B) 曼哈顿图。红线表示全基因组显著性阈值（$-log_{10}(5 \times 10^{-8})$）。

我们描述性地使用七种组织特异性GS注释（图4）和十种血液相关细胞类型特异性GSP注释（图5），研究了492，557个单核苷酸多态性的功能潜能。当GS和GSP评分截断值为0.5时，24%的单核苷酸多态性至少在七种组织类型中的一种组织中获得注释（图4A），15.4%的单核苷酸多态性至少在10种血液相关细胞类型特异性注释中的一种注释中获得注释（图5A）。获得注释的单核苷酸多态性比例从肺组织的5.66%到胃肠道组织的10.38%不等（图4B），从初始T $CD8^+$ 记忆细胞的3.43%到初始调节性T细胞的6.99%不等（图5B）。我们还使用对数优势比，测量了不同组织类型和细胞类型中获得注释的单核苷酸多态性之间的重叠程度（图4C和图5C）。与文献中关于肌肉和肺组织显示出更高表达数量性状位点共享水平的结论一致，

而血液显示出最低水平 [18, 33], 我们的研究结果表明，肌肉、肺和心脏组织中获得注释的单核苷酸多态性与其他组织类型的重叠更多，而血液组织中获得注释的单核苷酸多态性重叠更少（图4C）。最后，我们观察到不同类型的T细胞（初始辅助记忆、初始辅助性、效应/记忆富集型、调节型、$CD8^+$ 初始和$CD8^+$ 记忆T细胞）彼此之间的重叠更多，而中性粒细胞、初始B细胞和自然杀伤细胞之间的重叠较少（图5C）。

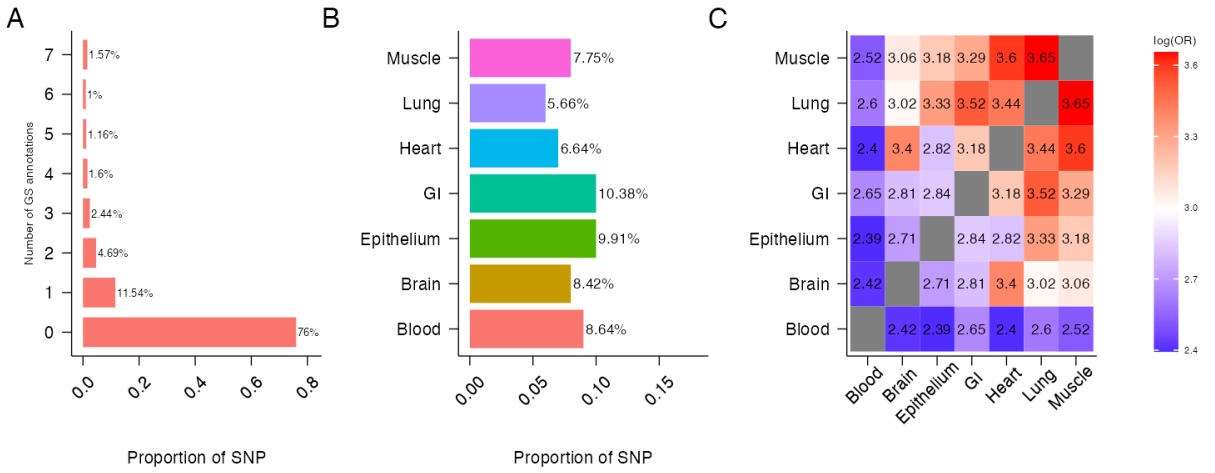

图4：整合七项GenoSkyline（GS）注释后492,557个单核苷酸多态性的特征。（A）单核苷酸多态性获得注释的GS组织数量。（B）每种GS组织类型中获得注释的单核苷酸多态性比例。（C）由七种GS组织类型注释的单核苷酸多态性之间的重叠，使用对数优势比计算。

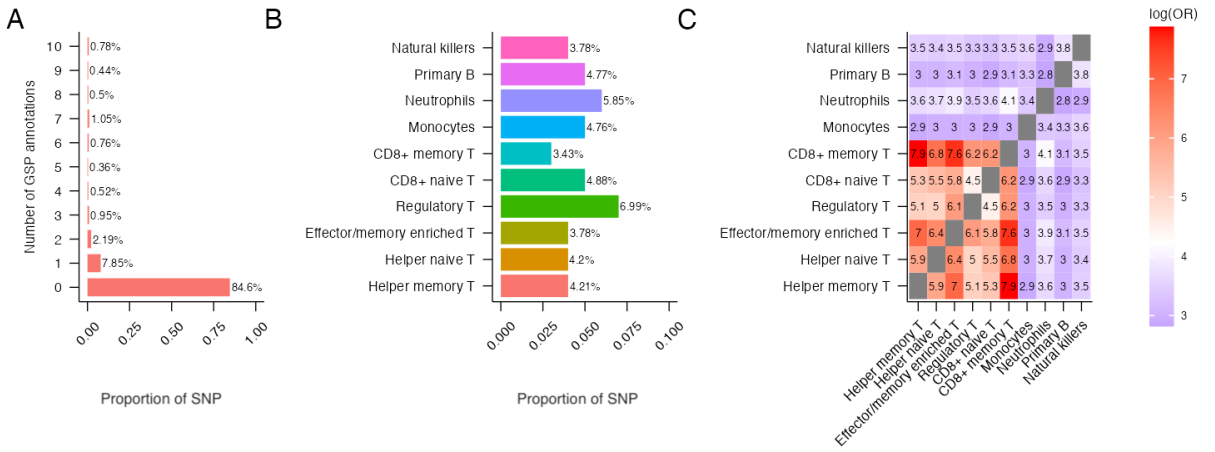

图5：整合10项血液相关GenoSkylinePlus（GSP）注释后492,557个单核苷酸多态性的特征。（A）单核苷酸多态性获得注释的GSP组织数量。（B）每项血液相关GSP注释中获得注释的单核苷酸多态性比例。（C）由10项血液相关GSP注释的单核苷酸多态性之间的重叠，使用对数优势比计算。

系统性红斑狼疮（SLE）与类风湿关节炎（RA）全基因组关联研究的整合

使用***GenoSkyline***（干燥综合征）注释进行组织层面的研究

我们将multi-GPA-Tree方法应用于SLE和RA全基因组关联研究以及组织特异性GS注释，以识别与SLE和RA分别呈边际关联及联合关联的单核苷酸多态性，并刻画与单一及多种性状风险相关单核苷酸多态性有关的功能注释。在名义全局错误发现率0.05水平下，multi-GPA-Tree识别出394个与SLE和RA均呈联合关联的单核苷酸多态性、1087个与SLE呈边际关联的单核苷酸多态性，以及791个与RA呈边际关联的单核苷酸多态性（表1）。

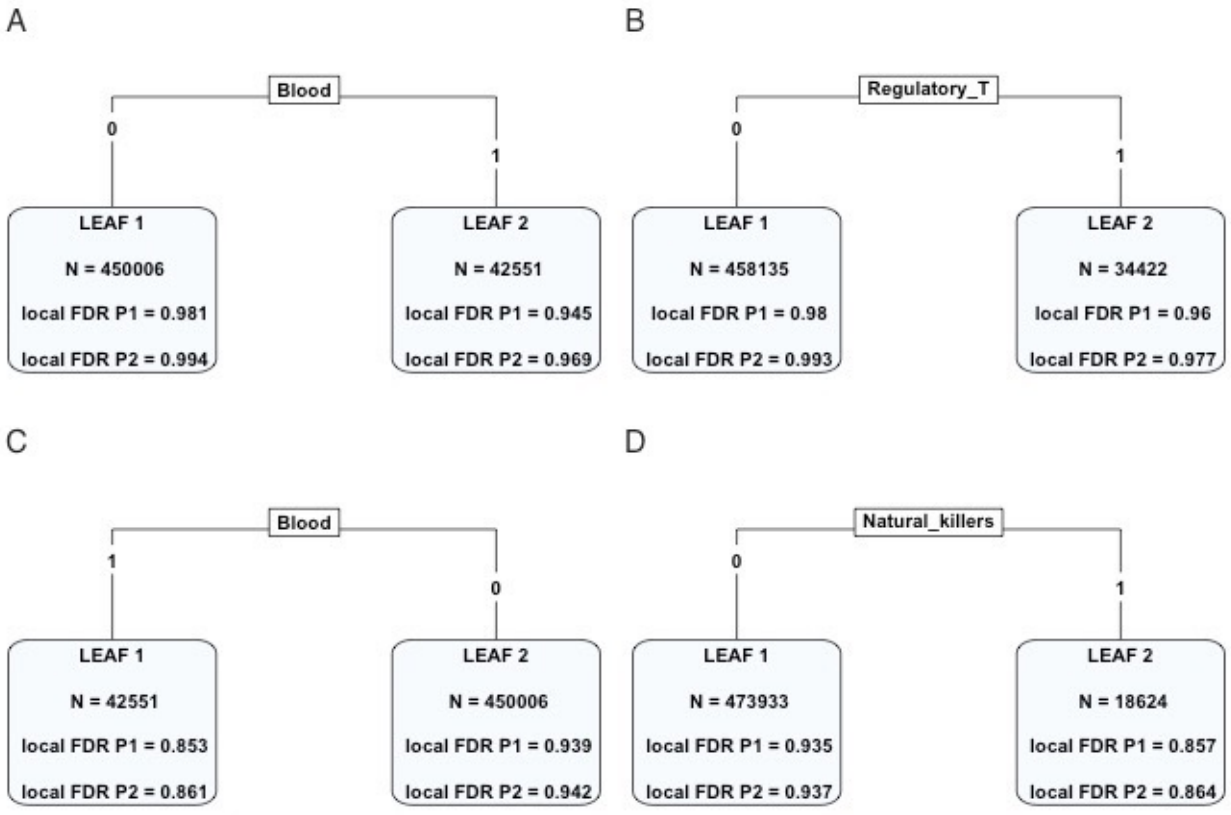

图6：**真实数据应用结果**。multi-GPA-Tree模型返回的树状结构：(A)整合SLE（P1）、RA（P2）和GenoSkyline注释；(B)整合SLE（P1）、RA（P2）和GenoSkylinePlus注释；(C)整合UC（P1）、CD（P2）和GenoSkyline注释；(D)整合UC（P1）、CD（P2）和GenoSkylinePlus注释。

在采用组织特异性GS注释对系统性红斑狼疮和类风湿关节炎进行联合分析时，原始multi-GPA-Tree模型将血液组织识别为根节点，并包含2个叶节点（图6A）。进一步研究发现，156个与系统性红斑狼疮和类风湿关节炎均相关的单核苷酸多态性、336个与系统性红斑狼疮边际相关的单核苷酸多态性，以及306个与类风湿关节炎边际相关的单核苷酸多态性，也都带有血液组织注释。在这156个同时相关且带有血液组织注释的单核苷酸多态性中，有118个属于蛋白质编码型，其中1号染色体、6号染色体、2号染色体和17号染色体上的蛋白质编码单核苷酸多态性数量最多。2号染色体上的 *PLCL1* 基因、10号染色体上的 *IL2RA* 基因和6号染色体上的 *UHRF1BP1* 基因所关联的蛋白质编码单核苷酸多态性数量最多：与编码 *PLCL1* 基因相关的单核苷酸多态性有5个，另有4

数据整合 tion	方法	# 边缘性地 相关 with P1 （系统性红斑狼疮/溃疡性结肠炎）	# 边缘性地 相关 with P2 (RA/CD)	# 共同 as- 相关 with P1 and P2 (系统性红斑狼疮+类风湿关节炎/ 溃疡性结肠炎+克罗恩病)	选定注释
系统性红斑狼疮+类风湿关节炎+干燥综合征	multi-GPA-Tree	1087	791	394	血液
系统性红斑狼疮+干燥综合征	GPA-Tree	696	-	-	血液
类风湿关节炎+干燥综合征	GPA-Tree	-	470	-	血液
系统性红斑狼疮+类风湿关节炎+GSP	multi-GPA-Tree	1065	760	383	调节性T细胞
系统性红斑狼疮+GSP	GPA-Tree	830	-	-	初始B细胞 调节性T细胞 记忆辅助性T细胞
类风湿关节炎+GSP	GPA-Tree	-	634	-	调节性T细胞 记忆辅助性T细胞 自然杀伤细胞
溃疡性结肠炎+克罗恩病+干燥综合征	multi-GPA-Tree	5430	5041	5041	血液
溃疡性结肠炎+干燥综合征	GPA-Tree	1566	-	-	血液
克罗恩病+干燥综合征	GPA-Tree	-	3185	-	血液
溃疡性结肠炎+克罗恩病+GSP	multi-GPA-Tree	4995	4912	4576	自然杀伤细胞
溃疡性结肠炎+GSP	GPA-Tree	1654	-	-	自然杀伤细胞 单核细胞 效应/记忆 富集型T细胞
克罗恩病+GSP	GPA-Tree	-	3232	-	调节性T细胞 自然杀伤细胞 单核细胞 效应/记忆 富集型T细胞 初始B细胞 调节性T细胞

表1：**真实数据应用结果汇总**。将系统性红斑狼疮（SLE）与类风湿关节炎（RA）的全基因组关联研究，以及克罗恩病（CD）与溃疡性结肠炎（UC）的全基因组关联研究，与GenoSkyline（干燥综合征）和GenoSkylinePlus（GSP）注释相结合，并使用multi-GPA-Tree和GPA-Tree方法进行联合分析和单独分析时，联合关联和边际关联的单核苷酸多态性数量。所有分析均包含492，557个单核苷酸多态性，以及7种组织特异性干燥综合征注释和10种血液相关细胞类型特异性GSP注释。推断基于在名义水平0.05下控制全局错误发现率。

个单核苷酸多态性分别与编码 *IL2RA* 基因和 *UHRF1BP1* 基因相关。已知 *PLCL1* 基因可通过调节NLRP3炎性小体促进炎症反应；NLRP3炎性小体是免疫系统的组成部分，与促炎细胞因子的激活和分泌有关 [35]。类似地，有研究报道 *IL2RA* 基因在活化的T细胞和B细胞、调节性T细胞、活化的单核细胞和自然杀伤细胞中表达 [36, 37], ；而 *UHRF1BP1* 基因参与非保守氨基酸改变，并与系统性红斑狼疮自身抗体靶向的RNA加工复合物有关 [38]。

我们还发现，染色体4、17、3、3，和16上分别有3个单核苷酸多态性已知可编码 *BANK1*、 *PGAP3*、*PLCL2*、 *RASA2*和 *TXNDC11* 基因。 *BANK1* 主要在CD19$^+$ B细胞中表达，是已知的SLE和RA易感基因 [39, 40]；动物研究显示， *PGAP3* 基因敲除与凋亡细胞清除减少相关，而凋亡细胞清除是自身免疫的一条因果通路 [41]；已知 *PLCL2*编码一种B细胞受体信号传导的负调节因子，对控制免疫反应十分重要，同时也是已知的RA易感基因 [42]。最后，尽管尚未在SLE和RA背景下进行研究，但已知 *RASA2* 变异与复合过敏性疾病相关 [43] ，且*TXNDC11* 已知参与甲状腺激素生物合成 [44]。

此外，我们通过将GS注释分别整合到SLE和RA的全基因组关联研究中，实现了GPA-Tree方法。对multi-GPA-Tree结果进行验证后发现，在SLE和RA各自的GPA-Tree分析中，血液组织均位于根节点。在单独GPA-Tree分析中，我们发现696个单核苷酸多态性与SLE相关，470个单核苷酸多态性与RA相关，其中分别有229个SLE相关和224个RA相关的单核苷酸多态性也被注释为与血液组织相关。在SLE和RA联合分析中鉴定出的排名前三的基因里，单性状分析也发现，一个或多个与 *IL2RA* 和 *PLCL1*基因相关的蛋白质编码单核苷酸多态性分别与SLE和RA有关。然而，单性状分析发现，与 *UHRF1BP1* 基因相关的蛋白质编码单核苷酸多态性与SLE有关，但与RA无关。

使用***GenoSkylinePlus***（***GSP***）注释进行细胞类型层面研究

根据观察到的血液组织GS注释与SLE和RA之间的关系，在分析的第二阶段，我们将multi-GPA-Tree 方法应用于SLE和RA全基因组关联研究以及10种血液相关的细胞类型特异性GSP注释，以识别与SLE和RA边际关联及联合关联的单核苷酸多态性，并描述与单一及多个性状的风险相关单核苷酸多态性有关的血液相关GSP功能注释。在名义全局错误发现率水平0.05下，multi-GPA-Tree识别出383个与SLE和RA联合关联的单核苷酸多态性、1,065个与SLE边际关联的单核苷酸多态性，以及760个与RA边际关联的单核苷酸多态性（表1）。联合分析还识别出位于根节点的初始调节性T细胞（图6B）；其中，95个与SLE和RA均联合关联、191个与SLE边际关联以及176个与RA边际关联的单核苷酸多态性也被注释为调节性T细胞。在这95个联合关联且也被注释为调节性T细胞的单核苷酸多态性中，69个为蛋白质编码单核苷酸多态性，其中1,6、16和2号染色体上的蛋白质编码单核苷酸多态性数量最多。2号染色体上的 *PLCL1*基因、10号染色体上的 *IL2RA*基因和16号染色体上的 *TXNDC11* 基因的蛋白质编码单核苷酸多态性数量最多，分别有3个不同的蛋白质编码单核苷酸多态性与这3个基因的编码相关。

使用GPA-Tree方法进行的单独分析发现，SLE中有初始B细胞、调节性T细胞和辅助记忆T细胞，其中初始B细胞位于根节点。类似地，我们发现RA中有调节性T细胞、辅助记忆T细胞和自然杀伤细胞，其中调节性T细胞位于根节点。在单独的GPA-Tree分析中，我们发现830个单核苷酸多态性与SLE相关，634个单核苷酸多态性与RA相关。在与SLE相关的单核苷酸多态性中，176个被注释为初始B细胞相关，122个被注释为调节性T细胞相关，43个被注释为辅助记忆T细胞相关。在与RA相关的单核苷酸多态性中，132个同时被注释为调节性T细胞和自然杀伤细胞相关，148个被注释为调节性T细胞相关但未被注释为自然杀伤细胞相关，32个被注释为自然杀伤细胞相关但未被注释为调节性T细胞相关，35个被注释为辅助记忆T细胞相关。在SLE、RA和GSP注释联合分析中确定的排名前3的基因中，有一个或多个与 *IL2RA* 基因相关的蛋白质编码单核苷酸多态性也在SLE和RA的单性状分析中被发现。然而，与 *PLCL1* 基因相关的蛋白质编码单核苷酸多态性仅在SLE中被发现，而与 *TXNDC11* 基因相关的蛋白质编码单核苷酸多态性仅在RA中被发现。

整合溃疡性结肠炎（UC）和克罗恩病（CD）的全基因组关联研究

使用 ***GenoSkyline***（干燥综合征）注释进行组织层面的研究

我们还将 multi-GPA-Tree 方法应用于溃疡性结肠炎和克罗恩病的全基因组关联研究，以及组织特异性干燥综合征注释，以识别与溃疡性结肠炎和克罗恩病边际相关及联合相关的单核苷酸多态性，并描述单一性状和多性状风险相关单核苷酸多态性的功能相关性。在名义全局错误发现率0.05水平下，multi-GPA-Tree识别出5,041个与溃疡性结肠炎和克罗恩病均联合相关的单核苷酸多态性，

5,430个与溃疡性结肠炎边际相关的单核苷酸多态性，以及5,041个与克罗恩病边际相关的单核苷酸多态性（表1）。在这项联合分析中，原始multi-GPA-Tree模型将血液组织确定为根节点，并包含2个叶节点（图6C）。进一步研究显示，1,319个与溃疡性结肠炎和克罗恩病均联合相关的单核苷酸多态性、1,453个与溃疡性结肠炎边际相关的单核苷酸多态性，以及1,319个与克罗恩病边际相关的单核苷酸多态性也被注释为与血液组织相关。在这1,319个联合相关且经血液组织注释的单核苷酸多态性中，有990个为蛋白质编码单核苷酸多态性。1号和2号染色体上的蛋白质编码单核苷酸多态性数量最多，其次是17号和5号染色体。2号染色体上的 *THADA* 和 *ATG16L1*基因、5号染色体上的 *C5orf56* 基因以及17号染色体上的 *IKZF3* 基因所含的蛋白质编码单核苷酸多态性数量最多；与编码 *THADA* 和 *IKZF3* 基因相关的单核苷酸多态性各有9个，与编码 *ATG16L1* 和 *C5orf56* 基因相关的单核苷酸多态性各有8个。尽管 *THADA* 基因并未直接涉及溃疡性结肠炎或克罗恩病的发病机制，但已知它会影响脂肪生成等代谢机制 [45]。相比之下， *ATG16L1* 基因的遗传变异是克罗恩病发病机制中研究最多的变异之一，在病原体清除、细胞因子产生、蛋白质调控和内质网应激反应调控中发挥作用 [46, 47]。类似地，在克罗恩病和溃疡性结肠炎患者中观察到 *IKZF3*表达增加；这种转录因子在调控B淋巴细胞增殖和分化方面发挥重要作用 [48, 49], ；而已知 *C5orf56* 会影响 *IRF1*的免疫刺激特异性增强子，该基因已被证实与克罗恩病的发病机制有关 [50, 51, 52]。

我们还发现，在染色体 5、6 和 9 上各有 7 个单核苷酸多态性，已知这些染色体编码 *FYB*、*BACH2* 和*DOCK8* 基因；在染色体 4 上，各有 6 个单核苷酸多态性，已知编码 *BANK1*、 *LEF1*和 *NFKB1* 基因。 *FYB* 基因与 T 细胞信号传导有关，在 IL-2A表达中发挥作用，并已知与某些自身免疫调控 [53, 54]相关。同样， *BACH2* 是对 B细胞调节活性、T细胞功能和分化至关重要的基因，也是克罗恩病和溃疡性结肠炎的已知易感位点 [55, 56]；已知 *DOCK8* 可调节多种免疫亚型，包括淋巴细胞，并参与免疫突触形成和病原体增殖 [57]；而 *NFKB1* 是免疫反应、细胞凋亡和细胞增殖的已知转录调节因子，在溃疡性结肠炎和克罗恩病患者中均呈上调表达 [58]。相反，尽管 *BANK1* 是一个已知与系统性红斑狼疮和类风湿关节炎相关的 B 细胞基因 [40, 39], ，但目前仅发表了少量将特定 *BANK1* 变异与克罗恩病联系起来的研究 [59, 60] ，其在克罗恩病和溃疡性结肠炎发病机制中的作用仍未得到充分研究。对于 *LEF1* 基因也是如此，它是 Wnt信号通路中的一种已知介质 [61]。

使用 GPA-Tree方法对 UC 和 CD 进行单性状分析时，我们为 UC 识别出血液、胃肠道和上皮组织，为 CD 识别出血液和上皮组织；两种性状的根节点均为血液组织。GPA-Tree 识别出 1,566 个与 UC 相关的单核苷酸多态性，以及 3,185 个与 CD 相关的单核苷酸多态性，其中分别有 540 个与 UC 相关、960 个与 CD 相关的单核苷酸多态性也被注释为与血液组织相关。在 UC、CD 和 GS 注释的联合分析中识别出的排名前 4 的基因里，有一个或多个与 *THADA*、 *IKZF3* 和 *C5orf56* 基因相关的单核苷酸多态性也在 UC 和 CD 的单性状分析中被识别。不过，与 *ATG16L1* 基因相关的单核苷酸多态性仅在 CD 中被识别。

使用 ***GenoSkylinePlus (GSP)*** 注释进行细胞类型层面研究

在分析的第二阶段，我们使用 multi-GPA-Tree 方法，将 UC 和 CD 的全基因组关联研究与 10 项血液相关细胞类型特异性 GSP 注释相结合。在名义全局错误发现率 0.05 水平下，multi-GPA-Tree 确定了 4,576 个与 UC 和 CD 共同相关的单核苷酸多态性、4,995 个与 UC 边际相关的单核苷酸多态性，以及 4,912 个与 CD 边际相关的单核苷酸多态性（表1）。原始 multi-GPA-Tree 模型拟合将初始自然杀伤细胞识别为根节点，并包含 2 个叶节点（图6D）。进一步研究发现，507 个与 UC 和 CD 均共同相关的单核苷酸多态性、579 个与 UC 边际相关的单核苷酸多态性，以及 554 个

与 CD 边际相关的单核苷酸多态性被注释为自然杀伤细胞。在 507 个共同相关且注释为自然杀伤细胞的单核苷酸多态性中，有 360 个属于蛋白质编码型。1号和2号染色体上的蛋白质编码单核苷酸多态性数量最多，其次是5号和17号染色体。5号染色体上的 *C5orf56* 和 *IRF1* 基因，以及10号染色体上的 *FAM53B* 基因，拥有数量最多的蛋白质编码单核苷酸多态性：与编码 *C5orf56* 基因相关的单核苷酸多态性有 8 个，与编码 *IRF1* 基因相关的有 5 个，与编码 *FAM53B* 基因相关的有 4 个。我们还发现了各有 3 个单核苷酸多态性，分别已知编码2号染色体上的 *ATG16L1*和 *THADA*基因、17号染色体上的 *IKZF3* 和 *PGAP3* 基因、9号染色体上的 *DOCK8* 基因、10号染色体上的 *TSPAN14* 基因，以及11号染色体上的 *ETS1* 基因。已知 *FAM53B* 与体液免疫应答、免疫效应过程调节及淋巴细胞活化调节相关 [62]； *PGAP3* 表达降低已知与凋亡细胞清除受损有关，并且在 CD 和 UC 患者中均有观察到 [49]； *TSPAN14* 在参与免疫和炎症的免疫细胞类型中表达，与巨噬细胞和中性粒细胞呈正相关，与 CD 中的 T细胞呈负相关8 [63]；最后，已知 UC 患者的肠上皮细胞中 *ETS1* 过度表达 [64], ，且在 CD 发病机制中， *ETS1* 也与瘘管形成有关；瘘管形成是破坏性炎症导致的上皮缺损 [65]。

使用GPA-Tree方法进行的单独分析发现，UC和克罗恩病（CD）均涉及初始自然杀伤细胞、单核细胞、效应/记忆富集T细胞和调节性T细胞，且自然杀伤细胞位于根节点。此外，CD还涉及初始B细胞。在单独GPA-Tree分析中，我们发现有1，654个单核苷酸多态性与UC相关，有3，232个单核苷酸多态性与CD相关。在与UC相关的单核苷酸多态性中，186个同时注释到自然杀伤细胞和效应/记忆富集T细胞，134个注释到自然杀伤细胞但未注释到效应/记忆富集T细胞，112个注释到单核细胞，127个注释到调节性T细胞。同样，在与CD相关的单核苷酸多态性中，278个同时注释到自然杀伤细胞和效应/记忆富集T细胞，211个注释到自然杀伤细胞但未注释到效应/记忆富集T细胞，181个注释到单核细胞，161个注释到调节性T细胞，132个注释到初始B细胞。在UC、CD和GSP注释的联合分析中发现的排名前三的基因中，一个或多个与 *C5orf56*、 *IRF1* 和 *FAM53B* 基因相关的蛋白质编码单核苷酸多态性，也在UC和CD的单性状分析中被发现。

讨论

过去20年来，研究人员开展了多项全基因组关联研究，成功鉴定出二十多万个与性状风险相关的遗传变异 [1]。新型统计方法日益复杂，能够进一步充分利用全基因组关联研究数据中的丰富信息，这有助于发现许多此前未知的遗传变异，也有望带来更多新发现。目前的研究发现对于确定治疗方案和推动新药发现至关重要 [66, 67, 68]。然而，在发现新变异的同时，我们仍需弥补一个关键认知空白：遗传变异通过哪些功能机制和通路影响性状。众所周知，复杂性状往往由多种功能机制共同导致，而这些机制可由多个功能注释来描述 [69, 70]。因此，识别与性状相关的功能注释组合，能够为性状病因学提供有价值的见解。然而，据我们所知，目前尚缺乏能够识别协同影响性状的功能注释组合的统计方法。我们提出本文讨论的 multi-GPA-Tree 方法，以弥补这一空白。

与现有方法相比，multi-GPA-Tree 方法的总体优势在于，它能够从一组注释中自动选择功能注释的组合，而不会

过度增加模型的复杂性，并可用于加深我们对与单个或多个性状相关的功能机制的理解。
multi-GPA-Tree 方法通过采用结合迭代过程（EM算法）和多变量决策树算法的层次架构来实现这一目标。在模拟研究中，在针对单个和多个性状识别性状风险相关变异方面，multi-GPA-Tree 方法在曲线下面积、统计功效和I类错误控制方面始终表现优于LPM方法；此外，它还以极高的准确性明确识别出了相关注释与噪声注释（图2）。

在真实数据应用中，联合分析两种性状时，multi-GPA-Tree 在识别两种性状的风险相关单核苷酸多态性方面效率更高，并验证了文献中已有的若干注释发现（表1，图6）。我们将 multi-GPA-Tree 的真实数据应用结果与我们近期发表的方法“GPA-Tree” [21], 进行了比较。GPA-Tree 是一种统计方法，不利用性状之间的多效性关系，而是优先筛选与单一性状边际相关的变异。比较结果表明，在联合分析两种性状时，multi-GPA-Tree 始终能为两种性状识别出更多边际风险相关变异。显然，GPA-Tree 识别出更多与特定性状相关的注释，而 multi-GPA-Tree 则识别出两种联合分析性状之间大体共有的注释。例如，在系统性红斑狼疮（SLE）与类风湿关节炎（RA），以及溃疡性结肠炎（UC）与克罗恩病（CD）的联合分析和单独分析中，血液组织均被识别出来。同样，在联合分析 SLE 和 RA 时，调节性T细胞被识别为相关注释；单独分析 SLE 和 RA 时，也发现了这一共同注释。类似地，在联合分析 UC 和 CD 时，自然杀伤细胞被识别为相关注释；单独分析 UC 和 CD 时，也均发现了这一共同注释。总体而言，这些结果与既往文献一致：研究表明，SLE、RA、UC 和 CD 等自身免疫性疾病与血液组织 [71, 72, 73], 有关，SLE 和 RA 与调节性T细胞 [74, 75, 76, 77, 78], 有关，而 UC 和 CD 与自然杀伤细胞 [79, 80, 81, 82]有关。此外，除识别出若干此前已被发现与其他过敏性疾病、甲状腺疾病或代谢性状相关的 SLE、RA、UC 和 CD 候选基因（$RASA2$, $TXNDC11$, $THADA$）外，我们还验证了既往研究发现的 $PLCL1$、$IL2RA$和 $UHRF1BP1$ 基因与 SLE 和 RA [83, 84, 85, 87, 86, 38, 88, 89], 的关联，以及 $ATG16L1$, $C5orf56$ 和$IKZF3$ 基因与 UC 和 CD [90, 91, 92, 50, 93, 51, 49, 94]的关联。

从统计建模的角度来看，multi-GPA-Tree采用了若干假设。首先，我们假设在给定遗传变异的功能信息后，这些变异条件独立，这大大简化了模型，并使参数估计值能够高效计算。尽管这一假设忽略了同一基因组区域内单核苷酸多态性之间的连锁不平衡（LD）结构，但它仍使我们能够保守地推断风险相关变异：通过适度控制I类错误率，我们也可能将彼此处于连锁不平衡状态的单核苷酸多态性识别为风险相关变异。其次，我们假设全基因组关联研究关联P值中的信号与单核苷酸多态性的功能潜能相关，因此，全基因组关联研究数据和注释数据中应存在一定的功能信号，multi-GPA-Tree方法才能有效运行。模拟结果表明，至少10%的变异应针对至少一个特征进行功能注释，才能获得有效的参数估计值，并在名义水平上控制I类错误率。

本文介绍了一种名为 multi-GPA-Tree 的新型统计方法，该方法能够利用多个全基因组关联研究数据之间的多效性关系，并以高效的方式整合全基因组关联研究数据与组织和细胞类型特异性功能注释数据。与一些要求个体水平基因型数据以及遵循特定分布假设的注释数据的现有方法相比，multi-GPA-Tree 仅需全基因组关联研究数据的汇总统计量和用于分析的二元注释数据。这些特点使 multi-GPA-Tree 成为一种颇具吸引力且有效的工具，可用于将全基因组关联研究数据与功能注释数据进行整合分析。尽管
multi-GPA-Tree 在统计性能方面取得了令人鼓舞的改进，但其生物学意义仍需通过独立研究进行重复验证，新发现的

变异也需要经过独立验证。multi-GPA-Tree 的两个局限性在于：它无法处理连续型或计数型注释数据，并且将 multi-GPA-Tree 扩展至两个以上性状时，计算上仍可能具有挑战性。解决多个全基因组关联研究与多种注释数据整合相关的问题，是我们未来工作的重要方向。

致谢

我们感谢南卡罗来纳医科大学的 Paula S. Ramos 博士、Andrew Lawson 博士和 Kelly J. Hunt 博士，以及辛辛那提大学的 Hang J. Kim 博士，就相关主题提出宝贵意见，并在完成本研究的过程中给予指导和支持。本研究获得美国国家普通医学科学研究所（R01 GM122078）、美国国家药物滥用研究所（U01 DA045300）、美国国家人类基因组研究所（R21 HG012482）、美国国家老龄研究所（U54 AG075931）以及 Pelotonia免疫肿瘤学研究所（PIIO）的资助。本文内容完全由作者负责，并不一定代表资助方的官方观点。

利益冲突

无。

参考文献

[1] Buniello A, MacArthur JA, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D. 已发表的全基因组关联研究、靶向芯片及汇总统计量的NHGRI-EBI全基因组关联研究目录，2019。《核酸研究》。2019年1月8日；47(D1)：D1005-12。[2] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH. 寻找复杂疾病中缺失的遗传力。《自然》。2009年10月；461(7265)：747-53。[3] Lee SH, Wray NR, Goddard ME, Visscher PM. 利用全基因组关联研究估算疾病中缺失的遗传力。《美国人类遗传学杂志》。2011年3月11日；88(3)：294-305。[4] Maher B. 个人基因组：缺失遗传力案例。《自然》。2008年11月6日；456(7218)：18-22。[5] Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC, Webb TR. 基于千人基因组计划数据的冠状动脉疾病全基因组关联荟萃分析。《Nature genetics》。2015年；47(10)：1121。[6]Price AL, Spencer CC, Donnelly P. 理解常见疾病遗传基础的进展与前景。《英国皇家学会学报B辑：生物科学》。2015年12月22日；282(1821)：20151684。[7] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V. 对111个人类参考表观基因组进行整合分析。《自然》。2015年2月；518(7539)：317-30。[8] Andreassen OA, Djurovic S, Thompson WK, Schork AJ, Kendler KS, O'Donovan MC, Rujescu D, Werge T, van de Bunt M, Morris AP, McCarthy MI. 利用与心血管疾病风险因素的多效性，提高精神分裂症相关常见变异的检出率

。《美国人类遗传学杂志》。2013年2月7日；92(2)：197-209。[9] Stearns FW. 多效性的一百年：回顾。《遗传学》。2010年11月1日；186(3)：767-73。[10] Chung D, Kim HJ, Zhao H. graph-GPA：一种用于GWAS结果优先排序并探究多效性架构的图模型。《PLoS computational biology》。2017年2月17日；13(2)：e1005388。[11] Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N. 美国国家生物技术信息中心基因型与表型数据库。《Nature genetics》。2007年10月；39(10)：1181-6。[12] Giral H, Landmesser U, Kratzer A. 探索未知领域：全基因组关联研究对非编码RNA的探索。《心血管医学前沿》。2018年12月17日；5：181。[13] Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJ, Shishkin AA, Hatan M. 对自身免疫性疾病因果变异进行遗传和表观遗传精细定位。《自然》。2015年2月；518(7539)：337-43。[14] Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A. 系统定位调控DNA中的常见疾病相关变异。《科学》。2012年9月7日；337(6099)：1190-5。[15] Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, Kelsoe JR, O'donovan MC, Furberg H, Tobacco and Genetics Consortium, Bipolar Disorder Psychiatric Genomics Consor- tium. 并非所有单核苷酸多态性都生而平等：全基因组关联研究揭示，功能注释的单核苷酸多态性富集呈现一致模式。《PLoS遗传学》。2013年4月25日；9(4)：e1003449。[16] Ming J, Dai M, Cai M, Wan X, Liu J, Yang C. LSMM：一种将功能注释与全基因组关联研究相整合的统计方法。《生物信息学》。2018年8月15日；34(16)：2788-96。[17] Zablocki RW, Schork AJ, Levine RA, Andreassen OA, Dale AM, Thompson WK. 用于全基因组关联研究的协变量调节局部错误发现率。《生物信息学》。2014年8月1日；30(15)：2098-104。[18] GTEx联盟, Ardlie KG, Deluca DS, Segr`e AV, Sullivan TJ, Young TR, Gelfand ET, Trow- bridge CA, Maller JB, Tukiainen T, Lek M. 基因型-组织表达（GTEx）试点分析：人类多组织基因调控。《科学》。2015年5月8日；348(6235)：648-60。[19] Chung D, Yang C, Li C, Gelernter J, Zhao H. GPA：一种整合多效性与注释、对GWAS结果进行优先排序的统计方法。《PLoS遗传学》。2014年11月13日；10(11)：e1004787。[20] Ming J, Wang T, Yang C. LPM：一种利用多个全基因组关联研究及功能注释的汇总统计量刻画复杂性状间关系的潜变量概率单位模型。《生物信息学》。2020年4月15日；36(8)：2506-14。[21] Khatiwada A, Wolf BJ, Yilmaz AS, Ramos PS, Pietrzak M, Lawson A, Hunt KJ, Kim HJ, Chung D. GPA-Tree：一种用于功能注释树引导的GWAS结果优先排序的统计方法。《生物信息学》。2022年2月15日；38(4)：1067-74。[22] De'Ath G. 多元回归树：一种模拟物种—环境关系的新技术。《生态学》。2002年4月；83(4)：1105-17。[23] Moon TK. 期望最大化算法。《IEEE信号处理杂志》。1996年11月；13(6)：47-60。[24] Newton MA, Noueiry A, Sarkar D, Ahlquist P. 使用半参数层级混合方法检测差异基因表达。《生物统计学》。2004年4月1日；5(2)：155-76。

[25] Gohlke JM, Thomas R, Zhang Y, Rosenstein MC, Davis AP, Murphy C, Becker KG, Mattingly CJ, Portier CJ. 复杂疾病的遗传和环境通路。BMC Systems Biology. 2009年12月;3(1):1-5.[26] Kim YA, Wuchty S, Przytycka TM. 识别复杂疾病中的因果基因和失调通路。PLoS computational biology. 2011 Mar 3;7(3):e1001095.[27] Jia X, Shi N, Feng Y, Li Y, Tan J, Xu F, Wang W, Sun C, Deng H, Yang Y, Shi X. 使用多变量统计分析，鉴定出与七种自身免疫性/ 自身炎症性疾病相关的67个多效性基因。Frontiers in Immunology. 2020 Feb 3;11:30.[28] Lee PH, Anttila V, Won H, Feng YC, Rosenthal J, Zhu Z, Tucker-Drob EM, Nivard MG, Grotzinger AD, Posthuma D, Wang MM. 八种精神障碍的基因组关系、新位点和多效性机制。Cell. 2019 Dec 12;179(7):1469-82[29] Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, Rudan I, McKeigue P, Wilson JF, Campbell H. 人类复杂疾病和性状中广泛存在的多效性。《美国人类遗传学杂志》. 2011 Nov 11;89(5):607-18.[30] Langefeld CD, Ainsworth HC, Graham DS, Kelly JA, Comeau ME, Marion MC, Howard TD, Ramos PS, Croker JA, Morris DL, Sandling JK. 系统性红斑狼疮的跨祖源定位与遗传负荷。Nature communications. 2017 Jul 17;8(1):1-8.[31] Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S, Graham RR. 类风湿关节炎遗传学对生物学研究和药物发现的贡献。《自然》. 2014 Feb;506(7488):376-81.[32] De Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, Jostins L, Rice DL, Gutierrez- Achury J, Ji SG, Heap G. 全基因组关联研究指出，多个整合素基因的免疫激活与炎症性肠病相关。Nature genetics. 2017 Feb;49(2):256-61.[33] Lu Q, Powles RL, Wang Q, He BJ, Zhao H. 人类基因组整合性组织特异性功能注释为许多复杂性状带来新见解，并改善全基因组关联研究中的信号优先级排序。PLoS genetics. 2016 Apr 8;12(4):e1005947.[34] Lu Q, Powles RL, Abdallah S, Ou D, Wang Q, Hu Y, Lu Y, Liu W, Li B, Mukherjee S, Crane PK. 对人类基因组进行系统的组织特异性功能注释，揭示了与迟发性阿尔茨海默病相关的免疫相关DNA元件。PLoS genetics. 2017 Jul 24;13(7):e1006933.[35] Luo S, Li XF, Yang YL, Song B, Wu S, Niu XN, Wu YY, Shi W, Huang C, Li J. PLCL1通过NLRP3炎性小体调节类风湿关节炎中成纤维样滑膜细胞的炎症。Advances in Rheumatology. 2022 Jul 22;62.[36] Carr EJ, Clatworthy MR, Lowe CE, Todd JA, Wong A, Vyse TJ, Kamesh L, Watts RA, Lyons PA, Smith KG. IL2RA与系统性红斑狼疮和ANCA相关性血管炎的遗传关联呈现相反特征。BMC Medical Genetics. 2009年12月;10(1):1-7.[37] Caruso C, Candore G, Cigna D, Colucci AT, Modica MA. 可溶性IL-2受体的生物学意义。Mediators of Inflammation. 1993 Jan 1;2(1):3-21.[38] Gateva V, Sandling JK, Hom G, Taylor KE, Chung SA, Sun X, Ortmann W, Kosoy R, Fer- reira RC, Nordmark G, Gunnarsson I. 一项大规模重复研究鉴定出TNIP1、PRDM1、JAZF1、UHRF1BP1和IL10为系统性红斑狼疮的风险位点。Nature genetics. 2009 Nov;41(11):1228-33.[39] Kozyrev SV, Abelson AK, Wojcik J, Zaghlool A, Reddy L, Prasad MV, Sanchez E, Gunnarsson I, Svenungsson E, Sturfelt G, J¨onsen A. B细胞基因BANK1中的功能变异与系统性红斑狼疮相关。Nature genetics. 2008 Feb;40(2):211-6.

[40] Orozco G, Abelson AK, Gonz´alez-Gay MA, Balsa A, Pascual-Salcedo D, Garc´ıa A, Fern´a ndez- Gutierrez B, Petersson I, Pons-Estel B, Eimon A, Paira S. BANK1基因功能变异研究：类风湿关节炎。《关节炎与风湿病》：美国风湿病学会官方期刊。2009 Feb;60(2):372-9.[41] Wang Y, Murakami Y, Yasui T, Wakana S, Kikutani H, Kinoshita T, Maeda Y. 脂筏中糖基磷脂酰肌醇锚定蛋白富集对自身免疫控制的意义。《生物化学杂志》。2013 Aug 30;288(35):25490-9.[42] Bowes J, Ho P, Flynn E, Ali F, Marzo-Ortega H, Coates LC, Warren RB, McManus R, Ryan AW, Kane D, Korendowych E. 在大型银屑病关节炎队列中全面评估类风湿关节炎易感位点。风湿病年鉴。2012 Aug 1;71(8):1350-4.[43] Ferreira MA, Vonk JM, Baurecht H, Marenholz I, Tian C, Hoffman JD, Helmer Q, Tillander A, Ullemar V, Van Dongen J, Lu Y. 哮喘、花粉症和湿疹的共同遗传起源揭示了过敏性疾病的生物学机制。Nature genetics. 2017 Dec;49(12):1752-7.[44] Jaeger M, Sloot YJ, Horst RT, Chu X, Koenen HJ, Koeken VA, Moorlag SJ, de Bree CJ, Mourits VP, Lemmers H, Dijkstra H. 促甲状腺激素和甲状腺素支持人体免疫稳态。免疫学。2021 Jun;163(2):155-68.[45] Pau CT, Mosbruger T, Saxena R, Welt CK. 多囊卵巢综合征中的表型和组织表达与遗传风险的关系。《公共科学图书馆·综合》。2017 Jan 9;12(1):e0168870.[46] Salem M, Ammitzboell M, Nys K, Seidelin JB, Nielsen OH. ATG16L1：克罗恩病中的多功能易感因子。自噬。2015 Apr 3;11(4):585-94.[47] Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, Albrecht M, Mayr G, De La Vega FM, Briggs J, G¨unther S. 一项非同义单核苷酸多态性的全基因组关联扫描鉴定出ATG16L1中一个克罗恩病易感变异。Nature genetics. 2007 Feb;39(2):207-11.[48] Huang C, Haritunians T, Okou DT, Cutler DJ, Zwick ME, Taylor KD, Datta LW, Maranville JC, Liu Z, Ellis S, Chopra P. 非裔美国人炎症性肠病易感性相关遗传位点的特征分析。《胃肠病学》。2015 Nov 1;149(6):1575-86.[49] S¨oderman J, Berglind L, Almer S. 基因表达-基因型分析提示GSDMA、GS-DMB和LRRC3C可能参与炎症性肠病易感性。《生物医学研究国际》。2015 Oct;2015.[50] Le´on AS, Bernstein CN, El-Gabalawy H, Eck P. IBD5位点变异使马尼托巴白人队列面临炎症性肠病风险。Clin Nutr. 2018;5:1-6.[51] Brandt M, Kim-Hellmuth S, Ziosi M, Gokden A, Wolman A, Lam N, Recinos Y, Daniloski Z, Morris JA, Hornung V, Schumacher J. 一种自身免疫性疾病风险变异：免疫刺激下由IRF1介导的反式主调控效应？《PLoS遗传学》。2021 Jul 27;17(7):e1009684.[52] Huff CD, Witherspoon DJ, Zhang Y, Gatenbee C, Denson LA, Kugathasan S, Hakonarson H, Whiting A, Davis CT, Wu W, Xing J. 克罗恩病与IBD5位点的遗传搭便车。分子生物学与进化。2012 Jan 1;29(1):101-11.[53] Azevedo Silva JD, Addobbati C, Sandrin-Garcia P, Crovella S. 系统性红斑狼疮：新旧易感基因与临床表现。《当代基因组学》。2014 Feb 1;15(1):52-65.[54] Addobbati C, Brand˜ao LA, Guimar˜aes RL, Pancotto JA, Donadi EA, Crovella S, Segat L, Sandrin- Garcia P. FYB基因多态性与系统性红斑狼疮（SLE）易感性相关。人类免疫学。2013 Aug 1;74(8):1009-14.

[55] Laffin MR, Fedorak RN, Wine E, Dicken B, Madsen KL. BACH2基因变异与克罗恩病术后复发相关。《美国外科医师学会杂志》。2018年5月1日；226(5):902-8。[56] Zhang B, Sun T. 调节溃疡性结肠炎发病机制的转录因子。《生物医学研究国际》。2020年8月24日；2020。[57] Kearney CJ, Randall KL, Oliaro J. DOCK8通过调节信号转导事件来控制免疫。《细胞与分子免疫学》。2017年5月；14(5):406-11。[58] Karban AS, Okazaki T, Panhuysen CI, Gallegos T, Potter JJ, Bailey-Wilson JE, Silverberg MS, Duerr RH, Cho JH, Gregersen PK, Wu Y. 对一种新型NFKB1启动子多态性进行功能注释，该多态性会增加溃疡性结肠炎风险。《人类分子遗传学》。2004年1月1日；13(1):35-45。[59] Li D, Achkar JP, Haritunians T, Jacobs JP, Hui KY, D'Amato M, Brand S, Radford-Smith G, Halfvarson J, Niess JH, Kugathasan S. SLC39A8中的一种多效性错义变异与克罗恩病及人类肠道微生物组组成相关。《胃肠病学》。2016年10月1日；151(4):724-32。[60] Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Philip Schumm L, Sharma Y, Anderson CA, Essers J. 宿主-微生物相互作用塑造了炎症性肠病的遗传架构。《自然》。2012年11月；491(7422):119-24。[61] Beisner J, Teltschik Z, Ostaff MJ, Tiemessen MM, Staal FJ, Wang G, Gersemann M, Perminow G, Vatn MH, Schwab M, Stange EF. TCF-1介导的Wnt信号通路调节潘氏细胞先天免疫防御效应分子HD-5和-6：对克罗恩病的启示。《美国生理学杂志-胃肠与肝脏生理学》。2014年9月1日；307(5):G487-98。[62] Xu Q, Chen S, Hu Y, Huang W. 有助于胰腺导管腺癌治疗的临床M2巨噬细胞相关基因。《国际癌细胞》。2021年12月；21(1):1-7。[63] Li Q, Gao X, Luo X, Wu Q, He J, Liu Y, Xue Y, Wu S, Rao F. 通过全血转录组分析鉴定与心源性栓塞性卒中免疫浸润相关的枢纽基因。《疾病标志物》。2022年1月15日；2022。[64] Li L, Miao X, Ni R, Miao X, Wang L, Gu X, Yan L, Tang Q, Zhang D. 上皮特异性ETS-1（ESE1ELF3）通过加速NF−κB激活调节溃疡性结肠炎中肠上皮细胞的细胞凋亡。《免疫学研究》。2015年6月；62(2):198-212。[65] Scharl M, Rogler G. 克罗恩病瘘管形成的病理生理学。《世界胃肠病理生理学杂志》。2014年8月8日；5(3):205。[66] Shu L, Blencowe M, Yang X. 将全基因组关联研究发现转化为冠状动脉疾病的新治疗靶点。《心血管医学前沿》。2018年5月30日；5:56。[67] Breen G, Li Q, Roth BL, O'donnell P, Didriksen M, Dolmetsch R, O'reilly PF, Gaspar HA, Manji H, Huebel C, Kelsoe JR. 将全基因组关联研究发现转化为精神病学的新疗法。《自然神经科学》。2016年11月；19(11):1392-6。[68] Visscher PM, Brown MA, McCarthy MI, Yang J. 全基因组关联研究发现的五年历程。《美国人类遗传学杂志》。2012年1月13日；90(1):7-24。[69] Petronis A. 表观遗传学是复杂性状和疾病病因学中的统一原则。《自然》。2010年6月；465(7299):721-7。[70] Zhang W, Voloudakis G, Rajagopal VM, Readhead B, Dudley JT, Schadt EE, Bj¨orkegren JL, Kim Y, Fullard JF, Hoffman GE, Roussos P. 综合转录组插补揭示了介导复杂性状易感性的组织特异性及共享生物学机制。Nature communications。2019年8月23日；10(1):1-3。

[71] Wahren-Herlenius M, D¨orner T. 系统性自身免疫性疾病的免疫致病机制。《柳叶刀》。2013年8月31日；382(9894):819-31。[72] Smith JB, Haynes MK. 类风湿关节炎——分子层面的认识。《内科学年鉴》。2002年6月18日；136(12):908-22。[73] Gleeson MH, Walker JS, Wentzel J, Chapman JA, Harris R. 克罗恩病和溃疡性结肠炎中的人类白细胞抗原。《Gut》。1972年6月1日；13(6):438-40。[74] Comte D, Karampetsou MP, Tsokos GC. T细胞作为系统性红斑狼疮的治疗靶点。《狼疮》。2015年4月；24(4-5):351-63。[75] Ohl K, Tenbrock K. 系统性红斑狼疮中的调节性T细胞。《欧洲免疫学杂志》。2015年2月；45(2):344-55。[76] Hoffman RW. T细胞在系统性红斑狼疮发病机制中的作用。《临床免疫学》。2004年10月1日；113(1):4-13。[77] Toh ML, Miossec P. T细胞在类风湿关节炎中的作用：新亚群和新靶点。《风湿病学当前观点》。2007年5月1日；19(3):284-8。[78] Leipe J, Skapenko A, Lipsky PE, Schulze-Koops H. 类风湿关节炎中的调节性T细胞。《关节炎研究与治疗》。2005年3月；7(3):1-7。[79] Manzano L, Alvarez-Mon M, Abreu L, Vargas JA, De la Morena E, Corugedo F, Dur`antez A. 活动性溃疡性结肠炎中自然杀伤细胞的功能障碍：白细胞介素-2可逆转自然杀伤活性的缺陷。《Gut》。1992年2月1日；33(2):246-51。[80] Takayama T, Kamada N, Chinen H, Okamoto S, Kitazume MT, Chang J, Matuzaki Y, Suzuki S, Sugita A, Koganei K, Hisamatsu T. 克罗恩病患者肠黏膜中的NKp44+ NKp46−和NKp44− NKp46+ 自然杀伤细胞失衡。《胃肠病学》。2010年9月1日；139(3):882-92。[81] Poggi A, Benelli R, Ven`e R, Costa D, Ferrari N, Tosetti F, Zocchi MR. 健康与疾病状态下的人类肠道相关自然杀伤细胞。Frontiers in Immunology。2019年5月3日；10:961。[82] Auer IO, Ziemer E, Sommer H. 克罗恩病中的免疫状态。V. 外周血体外自然杀伤细胞活性降低。《临床与实验免疫学》。1980年10月；42(1):41。[83] Ramos PS, Criswell LA, Moser KL, Comeau ME, Williams AH, Pajewski NM, Chung SA, Graham RR, Zidovetzki R, Kelly JA, Kaufman KM. 对系统性红斑狼疮（SLE）与16种自身免疫性疾病之间共享基因位点进行全面分析，发现遗传重叠有限。《PLoS遗传学》。2011年12月8日；7(12):e1002406。[84] Zhou R, Lin X, Li DY, Wang XF, Greenbaum J, Chen YC, Zeng CP, Lu JM, Ao ZX, Peng LP, Bai XC. 使用条件错误发现率方法鉴定骨质疏松症和/或类风湿关节炎的新遗传位点。《公共科学图书馆·综合》。2017年8月30日；12(8):e0183842。[85] Kasher M, Freidin MB, Williams FM, Cherny SS, Malkin I, Livshits G. 类风湿关节炎与不同骨质疏松表型之间共享的遗传架构。《骨与矿物质研究杂志》。2022年3月；37(3):440-53。[86] Hinks A, Ke X, Barton A, Eyre S, Bowes J, Worthington J, UK Rheumatoid Arthritis Genetics Con- sortium, 英国儿童与青少年风湿病学会研究组, Thompson SD, Langefeld CD, Glass DN. IL2RA\backslash CD25基因与幼年特发性关节炎的关联。《关节炎与风湿病》。2009年1月；60(1):251-7。[87] Carr EJ, Clatworthy MR, Lowe CE, Todd JA, Wong A, Vyse TJ, Kamesh L, Watts RA, Lyons PA, Smith KG. IL2RA与系统性红斑狼疮及ANCA相关性血管炎的遗传关联呈现不同模式。《BMC Medical Genetics》。2009年12月；10(1):1-7。

[88] Ramos PS、Shaftman SR、Ward RC、Langefeld CD。与SLE相关的基因是近期正选择的靶标。《自身免疫性疾病》。2014年10月；2014。[89] Delgado-Vega A、S´anchez E、L¨ofgren S、Castillejo-L´opez C、Alarc´on-Riquelme ME。系统性自身免疫性疾病遗传学的最新发现。《Current Opinion in Immunology》。2010年12月1日；22(6):698-705。[90] Lees CW、Barrett JC、Parkes M、Satsangi J。IBD遗传学新进展：与其他疾病共有的通路。《Gut》。2011年12月1日；60(12):1739-53。[91] Glas J、Konrad A、Schmechel S、Dambacher J、Seiderer J、Schroff F、Wetzke M、Roeske D、T¨or¨ok HP、Tonenchi L、Pfennig S。ATG16L1基因变异rs2241879和rs2241880（T300A）与德国人群克罗恩病易感性显著相关。《美国胃肠病学会官方期刊——ACG》。2008年3月1日；103(3):682-91。[92] Fisher SA、Tremelling M、Anderson CA、Gwilliam R、Bumpstead S、Prescott NJ、Nimmo ER、Massey D、Berzuini C、Johnson C、Barrett JC。溃疡性结肠炎的遗传决定因素包括ECM1基因座和克罗恩病相关的五个位点。《Nature Genetics》。2008年6月；40(6):710-2。[93] Cruz-Romero C、Guo A、Bradley WF、Vicentini JR、Yajnik V、Gee MS。克罗恩病患者的全基因组单核苷酸多态性与磁共振小肠造影特征之间的新关联。《Journal of Magnetic Resonance Imaging》。2021年1月；53(1):132-8。[94] Di Narzo AF、Peters LA、Argmann C、Stojmirovic A、Perrigoue J、Li K、Telesco S、Kidd B、Walker J、Dudley J、Cho J。抗TNF耐药克罗恩病队列的血液和肠道表达数量性状位点为炎症性肠病遗传关联位点提供信息。《Clinical and Translational Gastroenterology》。2016年6月；7(6):e177。