

基于基因表达量聚类的 eQTL 高效检测方法

吴泓杉^{1,2}, 赵亚宁^{1,2}, 顾林林^{1,2}, 方 铭^{1,2}

(1. 集美大学水产学院, 福建 厦门 361021; 2. 农业农村部东海海水健康养殖重点实验室, 福建 厦门 361021)

[摘要] 提出一个二倍体生物通用的 eQTL 鉴定的新方法。新方法利用一些基因组区域中基因表达量之间的相关性, 将具有较高相关性 ($r^2 > 0.5$) 的基因聚合在一起, 进而利用多性状联合分析模型鉴定出能同时调控多个基因表达的 cis-和 trans-eQTL。在对人类 GTEx 公共数据集进行分析时, 用新方法研究了 GTEx. v8 数据集的 6 种组织, 均发现了新的 eQTL, 使 eQTL 数量增加了 15.3%, 显著提高了 eQTL 的检测效力。

[关键词] eQTL; GTEx; 基因聚类; 多性状联合分析; 检测方法

[中图分类号] Q 811.4

An Efficient eQTL Detection Method Based on Gene Expression Clustering

WU Hongshan^{1,2}, ZHAO Yaning^{1,2}, GU Linlin^{1,2}, FANG Ming^{1,2}

(1. Fisheries College, Jimei University, Xiamen 361021, China;

2. Key Laboratory of Healthy Mariculture for the East China Sea, Ministry of Agriculture and Rural Affairs, Xiamen 361021, China)

Abstract: eQTLs are important tools for revealing the relationship between genetic variation and gene phenotypes. Improving eQTL detection efficiency helps in discovering new regulatory elements and target genes, which is crucial for understanding genetic regulatory mechanisms. This study developed a method based on gene expression clustering and multivariate analysis. By leveraging the correlations in gene expression levels within certain genomic regions, genes with high correlations ($r^2 > 0.5$) are grouped together. Subsequently, a multivariate analysis model is used to identify cis- and trans-eQTLs that regulate the expression of multiple genes simultaneously. When analyzing the human GTEx public dataset, the new method demonstrated excellent performance, significantly enhancing eQTL detection efficiency. We examined six tissues from the GTEx. v8 dataset and discovered new eQTLs, with an increase of 15.3% in their numbers. This study proposes a new eQTL identification method applicable to diploid organisms, providing a foundation for efficient eQTL mapping in diploid aquaculture species.

Keywords: eQTL; GTEx; gene clustering; multivariate analysis; detection method

0 引言

大量复杂性状的全基因组关联研究 (genome-wide association study, GWAS) 显示, GWAS 位点主要受调控元件控制, 而这些元件在 eQTL 中显著富集^[1-2]。识别各种组织中的 eQTL 具有重要意义, 因此已经进行了大量相关研究。举例来说, GTEx 数据库是 2020 年建立的一个庞大数据库, 包括 54

[收稿日期] 2024-08-14

[基金项目] 福建省自然科学基金项目 (2021J02045)

[通信作者] 方铭 (1979—), 博士, 教授, 博导, 从事鱼类遗传育种研究。E-mail: fangming618@126.com

http://xuebaobangong.jmu.edu.cn/zkb

种不同组织和 838 个样本^[3-4]。该数据库的创建初衷是通过收集大量人体组织的基因表达数据, 构建一个全面的遗传数据库, 以研究组织特异的基因表达和调控机制, 从而阐明遗传因素与复杂疾病和性状相关的分子机制, 提高人们对遗传调控机制的理解。

MatrIXeQTL^[5] 和 FastQTL^[6] 是两款 eQTL 分析的软件, 它们通过算法优化和计算速度提升, 显著改进了分析效率。这两个软件的基本原理都是基于单变量线性回归, 即每个基因的表达量被视为表型, 用于与基因组中的变异基因型进行关联分析。通常情况下, 基因表达量的获取需要使用微阵列芯片或转录组测序数据, 由于这些方法成本较高, 因此用于 eQTL 分析的个体数量或组织数量通常有限。较少的个体数量会降低 eQTL 的检测效率, 特别是对于 trans-eQTL 而言, 其较高的显著性阈值使得在基因组中寻找达到显著水平的位点变得更为困难^[7]。提高 eQTL 的检测效率, 发现更多的 eQTL, 这对深入理解复杂性状的遗传机制至关重要; 另外, 增强 eQTL 信号的强度有助于缩小 eQTL 的范围, 从而提高 eQTL 检测的准确性, 这有助于发现因果突变并阐明突变对靶基因表达的调控机制^[8]。

提高 eQTL 检测效率是广受关注的问题。充分利用多种信息来源是提高统计效率的主要手段。在 eQTL 检测方面, 联合多种组织的 eQTL 检测已经被广泛采用。通过不同组织基因表达量间的相关性, 采用多个组织联合鉴定能对多个组织发挥“一因多效”的 eQTL, 已被证实比单种组织分析具有更高的 eQTL 检测效率^[9]。在 GWAS 分析中, 通常存在多个性状的测量结果, 联合多个性状信息, 充分利用性状间相关信息进行 GWAS 分析, 也是一种提高 GWAS 位点检测效率的常用策略^[10]。在转录组测序数据中, 还存在另一种相关信息可以利用, 即基因表达量间相关信息。处于相邻基因组位置的基因往往受到相互元件调控而导致基因表达量之间存在关联, 这种现象在生物学上被称为“共调节”。但是, 尚未有利用该信息进行 eQTL 定位的研究。本研究根据基因表达量间的相关信息, 使用聚类方法将相关性较高的基因聚成一类, 再对这一类基因使用多性状分析模型, 检测同时控制多个基因表达的 eQTL。通过增加基因表达量之间的相关信息, 提高 eQTL 的检测效率。

1 材料与方法

1.1 材料

本研究使用的 GTEx. v8 数据集是从 dbGaP 网站获取的, dbGaP 登录号为 phs000424. v8. p2 (GTEx), 该数据集包含对 838 名捐献者的 54 种个体组织类型的全基因组测序 (whole genome sequencing, WGS) 数据和转录组测序 (RNA-seq) 数据^[2]。

数据集包含的组织类别过多, 在本方法的研究过程中, 只选取了部分样本量充足且研究价值较高的组织进行分析。本研究选取肝脏、脾脏、肌肉、胃、全血和肺这 6 种组织的全基因组测序数据和转录组测序数据, 其样本量分别为 226、241、803、359、755 和 578 个。

1.2 数据质量控制

1) 全基因组测序数据质量控制: 使用 Plink 软件^[11]对 GTEx. v8 数据集集中的 WGS 数据进行过滤, 将最小等位基因频率 (minor allele frequencies, MAF) 低于 0.05 的位点、SNP 缺失率高于 0.1 的位点和哈代温伯格平衡 (Hardy-Weinberg equilibrium, HWE) 低于 1×10^{-6} 的位点剔除掉。GTEx. v8 数据集集中的 WGS 数据共包括 66 463 168 个 SNP, 使用 Plink 软件过滤之后留下 8 379 990 个 SNP。

2) 转录组数据质量控制: 本研究的主要目标是 eQTL, 因此在分析之前要将基因表达量过低的基因剔除掉。用自己编写的 R 语言程序, 将基因表达量均值低于 3 的基因剔除掉, 最终在肝脏、脾脏、肌肉、胃、全血和肺等 6 种组织中, 分别留下 13 820、17 745、13 032、16 266、12 169 和 18 397 个表达量较高的基因。

1.3 方法

1.3.1 聚类分析

从 GTEx. v8 数据集获取的 RNA-seq 数据是 reads count, 即每个基因比对上的 reads 数量。reads count

数据易受转录本长度、测序总数等影响, 需要将其标准化之后用于后续分析。在本研究中将 reads count 转化成 RPKM (reads per kilobase of exon model per million mapped reads)。具体的转化公式如下:

$$N_{\text{RPKM}} = (N_{\text{EMR}} \times 10^9) / (R_{\text{T}} \times L)。$$
(1)

其中: N_{EMR} 是指 map 到基因上的 read 数 (exon mapped reads, EMR); R_{T} 是指 map 到基因组上的所有 read 数 (total mapped reads, TMP); L 是指 RNA 长度 (exon length, EL); N_{RPKM} 是指每百万 reads 中来自某基因每千碱基长度的 reads (RPKM) 数。

在对基因表达量聚类之前, 先使用 GTE_x. v8 数据集自带的协变量文件对基因表达量进行校正, 消除独立变量的影响, 以减小实验误差。再使用对数校正, 即将基因表达量 N_{RPKM} 转化成 $\log_2(N_{\text{RPKM}} + 1)$, 有利于线性回归分析。主要是通过 R 语言的 lm 函数和 predict 函数实现, 使用的公式为:

$$\begin{aligned} y &= \mu + \sum_{i=1}^n (b_i \times C_i) + e; \\ \hat{y} &= \hat{\mu} + \sum_{i=1}^n (\hat{b}_i \times C_i); \\ y_{\text{new}} &= y - \hat{y}。 \end{aligned}$$
(2)

其中: y 是指校正前的基因表达量; μ 是 y 的均值; b_i 是协变量的效应值; C_i 是协变量; e 是随机误差, $\hat{y}$ 、 $\hat{\mu}$ 和 $\hat{b}_i$ 均是预测值; y_{new} 是加入协变量校正后的基因表达量。

本研究根据基因表达量计算基因之间的皮尔逊相关系数, 度量两个基因之间的相关性。在自然科学领域中, 皮尔逊相关系数广泛用于度量两个变量之间的相关程度, 具体公式为:

$$p = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}。$$
(3)

其中: x_i 和 y_i 表示两个基因的表达式; $\bar{x}$ 和 $\bar{y}$ 表示两个基因表达量的平均值; p 则是两个基因之间的相关系数。

计算得到基因间的相关系数后, 使用类平均法 (group average method) 对基因进行聚类分析, 将相关系数在 0.5 以上的基因聚成一类。类平均法是分别计算一个类中的每个样本点与另一个类中的每个样本点之间的距离, 再计算一下平均值, 得到的均值就是两个类之间的距离, 具体公式为:

$$D(G_1, G_2) = \frac{1}{n_1 n_2} \sum_{x_i \in G_1} \sum_{x_j \in G_2} d(x_i, x_j)。$$
(4)

其中: n_1 和 n_2 分别表示两个类中的样本数; d 表示一个类中的每个样本点与另一个类中的每个样本点之间的距离; D 表示两类之间的距离; G_1 和 G_2 表示两个类。

1.3.2 使用 GEMMA 软件进行验证

GEMMA 软件的多元线性混合模型^[12]可以用于 Meta 分析, 但是得到的结果只有一个, 即聚在同一类的基因的 Meta 分析结果。因此, 在这里仅将其用来验证 MTAG 软件的结果。使用 GEMMA 软件进行 Meta 分析的基本原理为:

$$Y = WA + x\beta T + U + E。$$
(5)

其中: Y 表示 n 个样本、 d 个基因的基因表达量矩阵; W 是协变量矩阵; A 是包含相应系数的矩阵; x 是标记基因型; β 是 d 个基因的标记效应值; U 是取自亲缘关系矩阵的随机效应矩阵; E 是取自单位矩阵的误差矩阵。

GEMMA 软件可供选择的检测方式 (frequentist test) 共有三种, 分别是沃尔德检验 (Wald's test)、似然比检验 (likelihood ratio test) 和分值检验 (score test)。本研究选用似然比检验对 Meta 分析的结果进行验证, 具体步骤为: 1) 选取通过 MTAG 软件检测出的新 eQTL 聚成一类的基因; 2) 使用 GEMMA 软件对这一类的所有基因进行 Meta 分析; 3) 使用 GEMMA 软件对这一类基因中除了有新

检测出 eQTL 的基因之外的所有基因进行 Meta 分析；4）通过使用贝叶斯因子（Bayes factor，BF）检验，将步骤 2）结果中自然对数似然函数值和步骤 3）结果中自然对数似然函数值进行比较，判断 Meta 分析的结果是否可靠。贝叶斯因子检验的基本原理为：

$$2\ln(F_{12}) = 2\ln \frac{\text{lik}_1(X_n|\vartheta_{\text{MLE}})}{\text{lik}_2(X_n|\vartheta_{\text{MLE}})} - (d_1 - d_2)\ln(n)。$$

(6)

其中： F_{12} 表示贝叶斯因子； $\text{lik}_1(X_n|\vartheta_{\text{MLE}})$ 表示步骤 2）中得到的似然函数值； $\text{lik}_2(X_n|\vartheta_{\text{MLE}})$ 表示步骤 3）中得到的似然函数值，即基因类中所有基因 Meta 分析后的似然函数值； d_1 表示步骤 2）中的基因数量； d_2 表示步骤 3）中的基因数量； n 表示样本数量。

2 结果

本文研究了 GTEx 数据库中的 6 种组织，包括肝脏、胃、肌肉、肺、全血和脾脏的基因表达量。根据基因表达量的相关程度（ r^2 ）对基因进行了聚类。在这 6 个组织的聚类结果中，分别随机抽取了 35、26、17、22、32 和 18 个基因类进行进一步分析，总共涵盖了 150 个基因类，共包含 502 个基因。

2.1 发现新的 cis-eQTL

以胃组织中的一个包含 3 个基因的基因类作为示例。这个基因类中的 3 个基因分别是 ENSG00000276691、ENSG00000276550 和 ENSG00000272977，它们分别位于 Chr12：47788426—47788971、Chr15：22495570—22590815 和 Chr22：25476218—25479971。它们之间的基因相关系数分别为 0.509、0.529 和 0.540，均超过 0.5（见图 1）。

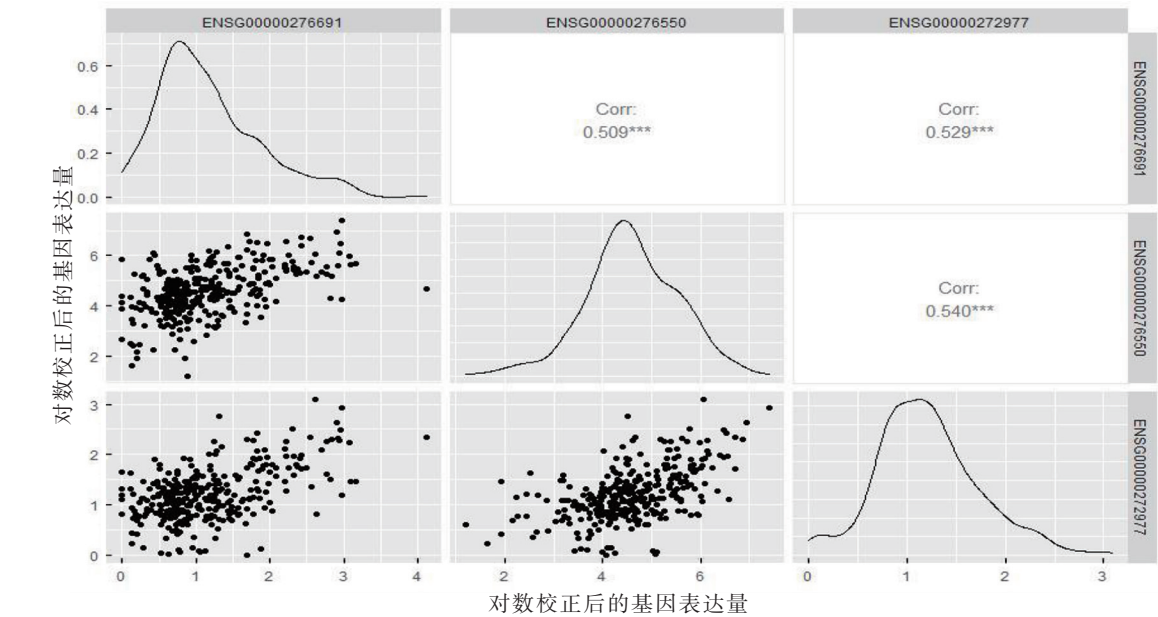


图 1 发现新的 cis-eQTL 的基因类中 3 个基因相关系数图

Fig.1 Plot of correlation coefficients of three genes in gene clusters where a new cis-eQTL was found

使用 GEMMA 软件对这 3 个基因分别进行 eQTL 分析。在 ENSG00000276691 基因中，未检测到显著的 eQTL（如图 2a 所示），而在另外两个基因中，分别检测到了显著的 eQTL。在 ENSG00000276550 基因上发现了一个位于 Chr15：22590507 的显著 eQTL（ $P=2.79 \times 10^{-12}$ ），而在 ENSG00000272977 基因上发现了一个位于 Chr22：25478497 的显著 eQTL（ $P=4.07 \times 10^{-13}$ ）（如图 2b、c 所示）。根据基因注释信息以及显著性 eQTL 在染色体上的位置，推测在 ENSG00000276550 基因上检测到的显著性 eQTL 距离基因左侧 95 kb，可能发挥着顺式调控作用；而在 ENSG00000272977 基因上检测到的显著性 eQTL 距离基因左侧 2 kb，同样可能具有顺式调控作用。

在对这个基因类使用 MTAG 软件进行 Meta 分析后, 发现 ENSG00000276550 基因上的 eQTL 信号明显提升 ($P=3.99 \times 10^{-18}$), 由 11 增加到 17; 而 ENSG00000272977 基因上的 eQTL 信号也明显提升 ($P=3.84 \times 10^{-21}$), 由 12 增加到 20 (如图 2e, 图 2f 所示)。在 ENSG00000215790 基因上原本未检测到显著性 eQTL (如图 2a 所示), 但经过 MTAG 软件进行 Meta 分析后, 在 ENSG00000276691 基因上的 Chr12: 47787477 处检测出了显著性 eQTL ($P=1.30 \times 10^{-10}$), 根据基因注释信息以及这些显著性 eQTL 在染色体上的位置, 推测新检测到的显著性 eQTL 距离基因左侧约 1 kb, 可能发挥顺式调控作用 (如图 2d 所示), 这是通过本研究方法发现的新的 cis-eQTL, eQTL 信号由 5 提升到 10。

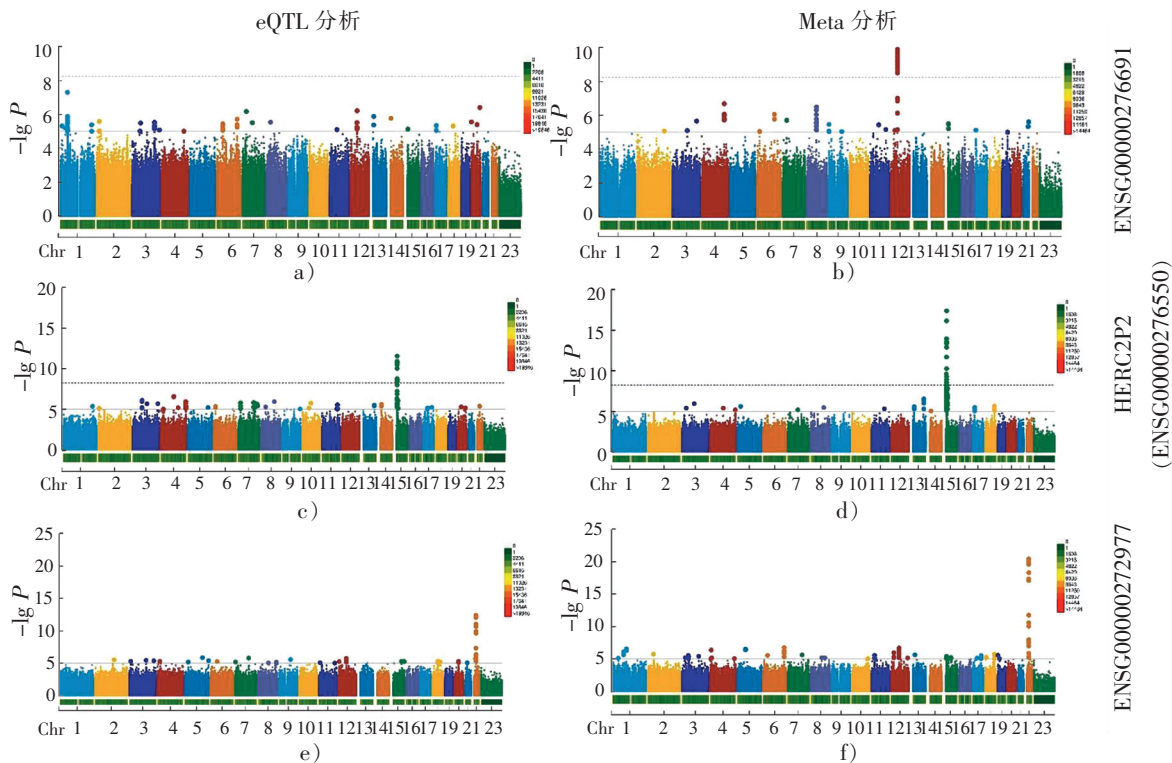


图 2 三个基因 eQTL 分析和 MTAG 分析曼哈顿图
Fig.2 The Manhattan map of eQTL analysis and MTAG analysis on three genes

为了验证结果的可靠性, 采用了 GEMMA 软件的多元线性混合模型进行验证。首先, 对已检测出显著性 cis-eQTL 的基因 ENSG00000276550 和 ENSG00000272977 进行 Meta 分析, 发现只有两个已知的 cis-eQTL, 分别位于 Chr15: 22590507 和 Chr22: 25478497 (如图 3 所示), 并未发现新的 eQTL 信号。接着对该基因类内的三个基因进行联合分析, 结果发现了新的显著 eQTL 信号, 就位于 Chr12: 47788289 (如图 4 所示)。

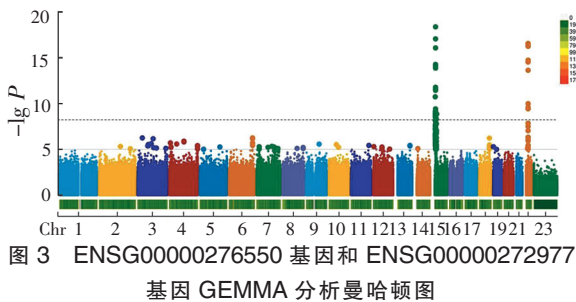


图 3 ENSG00000276550 基因和 ENSG00000272977 基因 GEMMA 分析曼哈顿图
Fig.3 Manhattan plot of GEMMA of ENSG00000276550 gene and ENSG00000272977 gene

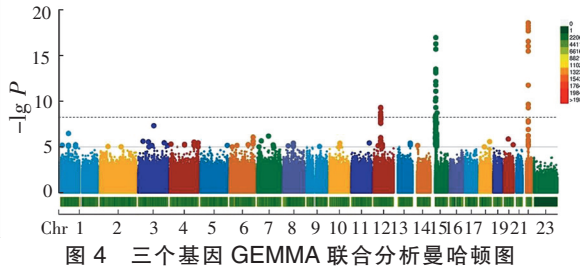


图 4 三个基因 GEMMA 联合分析曼哈顿图
Fig.4 The Manhattan map of GEMMA analysis on three genes

这个新发现的 eQTL 区域与用 MTAG 发现的区域吻合，同时也再次确认了 MTAG 检测到 ENSG00000276691 基因的新 eQTL 信号，由此，初步判断新发现的 cis-eQTL 是可靠的。再通过这两个方法得到的似然比结果计算贝叶斯因子，得到的贝叶斯因子值为 169.364。根据 Jeffreys 的标准^[13]，当贝叶斯因子大于 100 时，说明有极强的证据支持，因此判断新发现的 cis-eQTL 是可以信任的，即通过 MTAG 软件检测出的新 eQTL 是可靠的。通过 GEMMA 软件和 MTAG 软件的交叉验证，进一步说明了本方法的科学性和可靠性。

2.2 发现新的 trans-eQTL

本文以胃组织中的一个包含两个基因的基因类作为示例。这个基因类中的 2 个基因分别是 ENSG00000090104 和 ENSG00000110848，分别位于 Chr1: 192575727—192580031、Chr12: 9752486—9760901，它们之间的基因表达量相关系数为 0.73。值得注意的是，每个基因的基因表达量 ($\log_2(N_{\text{RPKM}} + 1)$) 近似服从正态分布（如图 5 所示）。

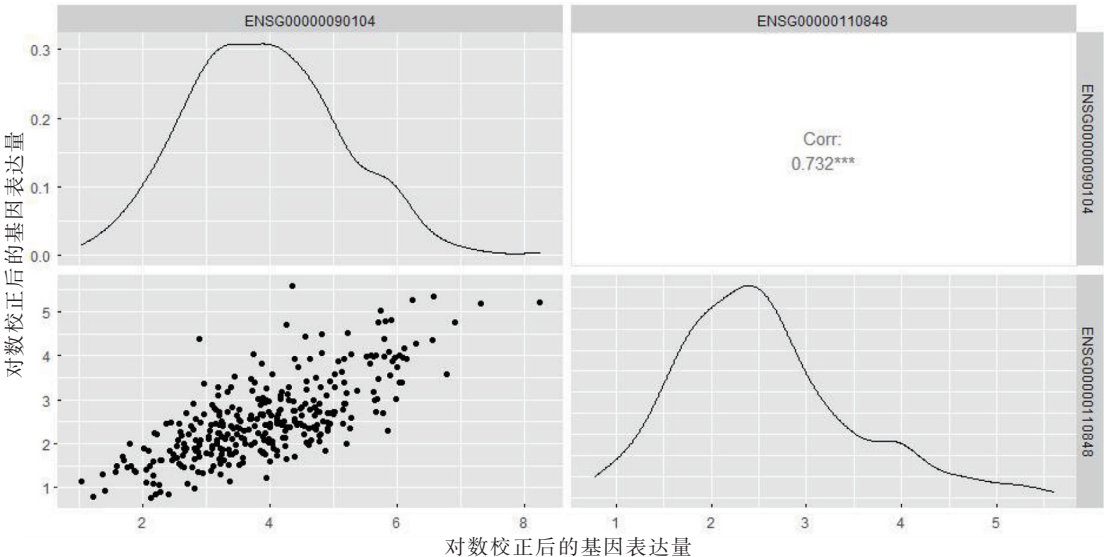


图 5 发现新的 trans-eQTL 的基因类中 2 个基因相关系数图

Fig.5 Plot of correlation coefficients of two genes in gene clusters where a new trans-eQTL was found
首先使用 GEMMA 软件对这个基因类中的 2 个基因进行了 eQTL 分析，但在这两个基因上均未检测到显著的 eQTL 信号（如图 6a、6b 所示）。

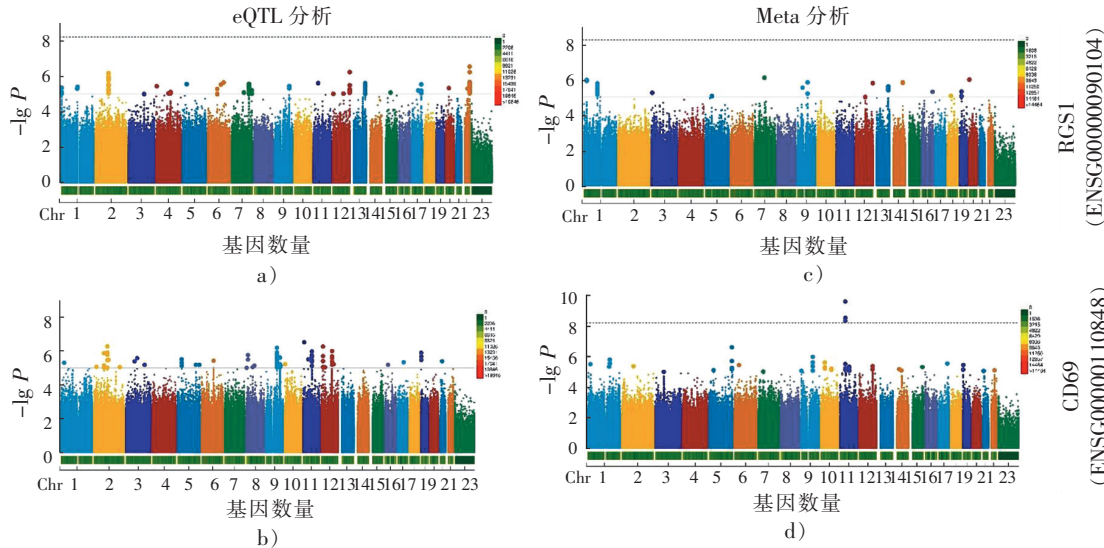


图 6 两个基因 eQTL 分析和 MTAG 分析曼哈顿图

Fig.6 The Manhattan map of eQTL analysis and MTAG analysis on two genes

然而, 在使用 MTAG 软件对这两个基因进行 Meta 分析后, 成功地在 ENSG00000110848 基因的 Chr11: 40869435 位置检测到了显著 eQTL 信号 ($P = 3.07 \times 10^{-6}$)。根据基因注释信息以及 eQTL 在染色体上的位置, 判断出这个新检测到的 eQTL 位于 11 号染色体上, 而 ENSG00000110848 基因位于 12 号染色体上, 因此推测这个新 eQTL 可能具有反式调控作用 (如图 6d 所示)。这个 eQTL 是通过本方法找到的新的 trans-eQTL。值得注意的是, 使用 GEMMA 软件的多元线性混合模型对这两个基因进行 Meta 分析时, 也在 Chr11: 40869435 位置发现了相同的新 trans-eQTL 信号, 与 MTAG 方法发现的 eQTL 信号一致 (如图 7 所示)。

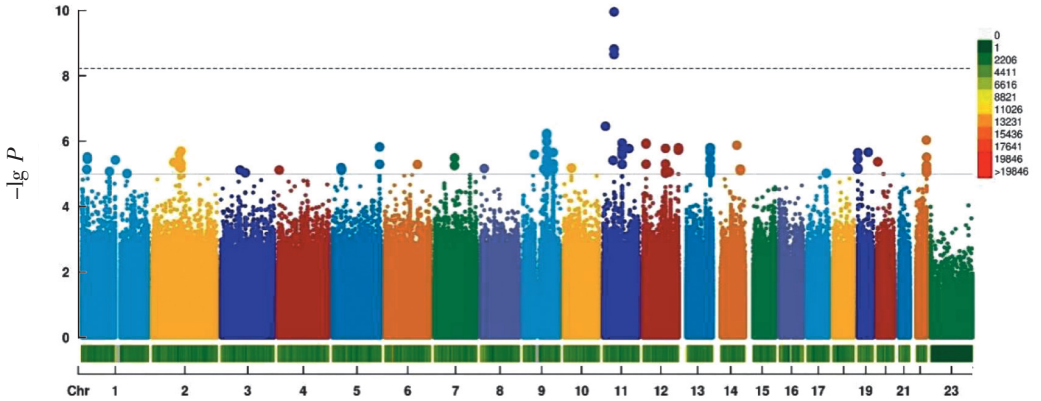


图 7 两个基因 GEMMA 分析曼哈顿图

Fig.7 The Manhattan map of GEMMA analysis on two genes

2.3 检测效率

为了评估多个基因联合分析的检测效率, 随机抽取了 502 个基因, 包括全血组织 102 个、肺组织 72 个、肝脏组织 118 个、肌肉组织 61 个、脾脏组织 67 个、胃组织 82 个, 再根据这 502 个基因的表达量进行了聚类, 最终形成了 150 个基因类。本研究首先利用 GEMMA 软件对这 502 个基因进行了基因组 eQTL 检测, 然后使用 MTAG 工具, 利用 GEMMA 获得的结果对 150 个基因类分别进行 Meta 分析。大约 6.7% (10 个) 基因类中检测到了新 eQTL (如图 8、图 9 所示), 而大约 81.5% (53 个) 基因类中发现了对已知 eQTL 信号的增强效应 (如图 10 所示)。

总共有 73 个基因在 eQTL 分析中被发现具有显著性 eQTL, 且分布在 65 个不同的基因类中。经 Meta 分析后, 在之前未发现显著性 eQTL 的 10 个基因上找到了新的 eQTL, 新增率 (新检测到的 eQTL 数量与之前已存在的显著性 eQTL 数量的比率) 达到了 15.3%。值得注意的是, 通过使用 GEMMA 软件的多元线性混合模型验证 MTAG 发现的结果, 同样也检测出了 10 个新的 eQTL, 因此可以确认本方法是可靠的。

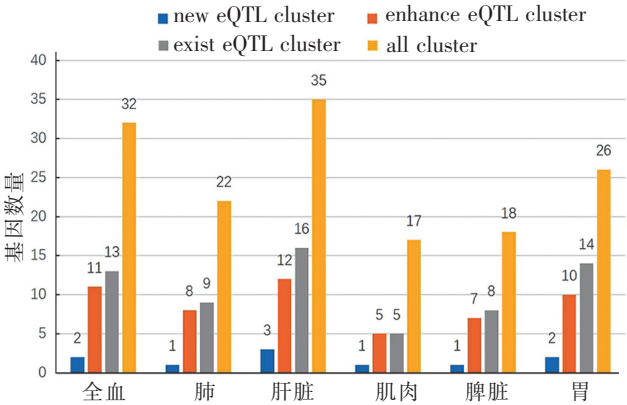


图 8 六种组织中基因类的数量

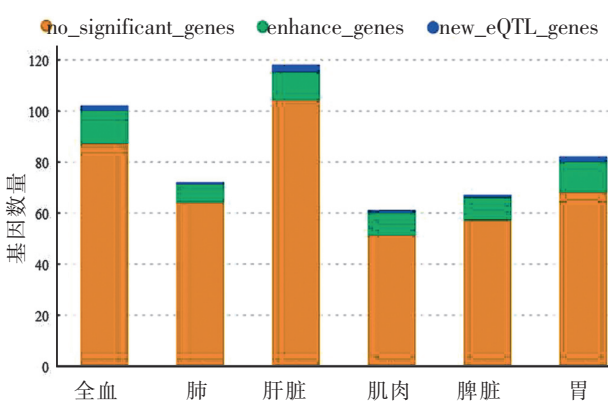
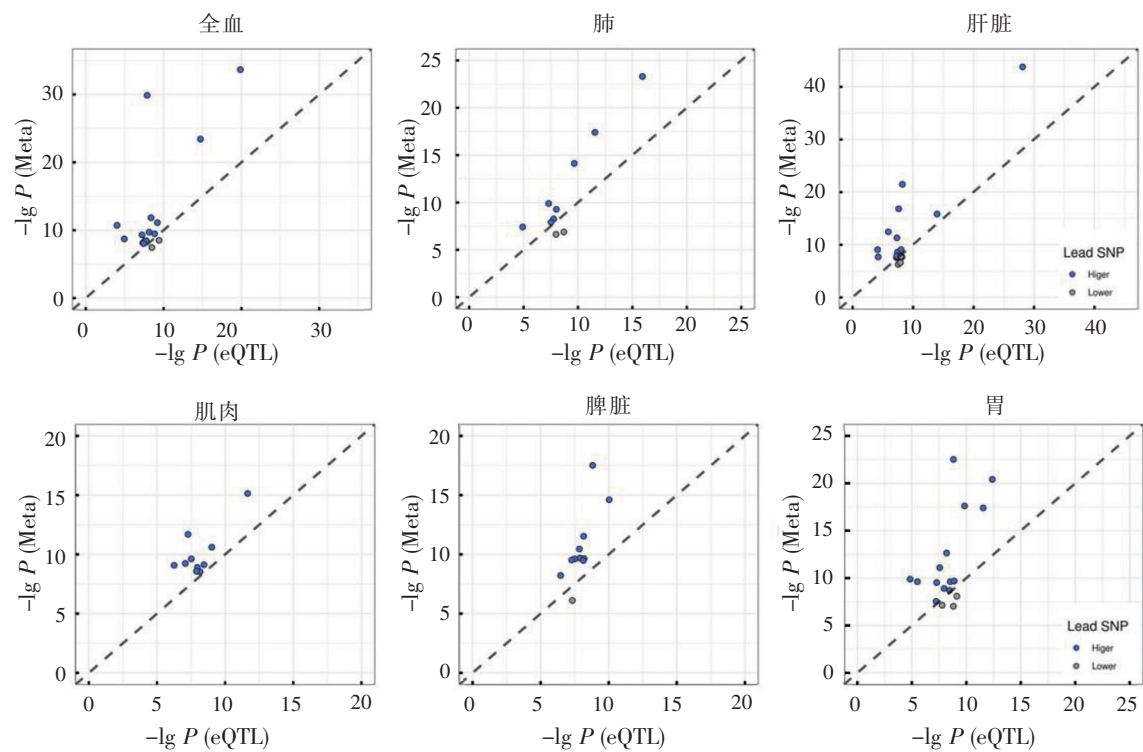


图 9 六种组织中基因的数量

Fig.8 The number of gene clusters in each of the six tissues Fig.9 The number of genes in each of the six tissues



蓝色的点表示 Meta 分析后 eQTL 信号提升; 灰色的点表示 Meta 分析后 eQTL 信号降低。

图 10 六种组织中增强的基因类信号强度变化图

Fig.10 Signal intensity variations of enhanced gene classes in six tissues

综上所述，本方法在人类的多个组织中表现出良好的性能，可适用于大多数组织。

3 讨论

本文选择了 GTEx. v8 数据集中的 6 种组织的全基因组测序数据和转录组数据，经过聚类分析和 Meta 分析，成功发现了每个组织中的新 eQTL。研究结果表明，这一方法可以明显提高检测效率。在研究过程中，使用本方法对 150 个基因类进行分析，成功检测出了 10 个新的显著性 eQTL，这表明本方法具有一定的应用潜力。由于不同组织间的基因表达量存在差异，eQTL 分析通常需要大量的样本，这会带来高昂的成本。然而，通过本研究开发的方法，可以在成本不变的情况下，挖掘出更多的显著性 eQTL，极大提高了 eQTL 分析的效率。值得一提的是，本方法适用于检测新的 cis-eQTL 和 trans-eQTL，能够满足 eQTL 分析中的大部分需求。

21 世纪 10 年代，就已经开始进行多组织 eQTL 研究。与单个组织的 eQTL 分析相比，使用这种统计框架进行多组织 eQTL 分析，可以明显提高在所有组织中都存在的 eQTL 的数量和显著性。此外，Sul 等^[14]开发了 MetaTissue (MT) 软件，该软件采用混合效应 Meta 分析框架，不仅解释了组织间基因表达的相关性，还考虑了跨组织效应的异质性。Flutre 等^[15]提出了一种贝叶斯层次模型 (eQTL-BMA)，该模型模拟了基因在多个组织中的联合分布，并利用基因间的信息来估计组织间 eQTL 共享模式的相对频率。此外，Allen 等^[16]开发了 R 语言 JAGUAR 软件包，该方法的特点在于将组织特异性视为随机效应，从而减少了需要测试的参数数量，提高了分析能力。这些研究为 eQTL 的鉴定做出了重要贡献。从生物学角度来看，SNP 和基因之间的关系很难用线性模型来解释，因为基因在不同组织中的表达水平存在差异。并且，通常情况下难以对每个组织都进行分析，因此可能会导致未能发现某些显著的 eQTL^[17]。本研究采用了一种新的方法，首先根据基因表达量将基因进行聚类，然后对同一类别内的不同基因进行 Meta 分析。与现有研究不同的是，本研究关注于同一组织内不同基因之间的

相互关系。当然,对于这些基因,有一个关键要求,即它们的表达量相关系数应高于0.5。

本方法的主要不足在于,在聚类过程中,某些基因类中的基因数量过多,会在后续的Meta分析中消耗大量计算资源和时间,成为方法计算速度的主要限制因素。如果研究人员感兴趣的基因恰好位于一个基因数量庞大的基因类中,他们可以通过提高聚类标准来减少基因数量。未来的改进,主要着眼于提高计算速度,并优化整个分析流程。

由于本方法在GTE_x数据集上展现了良好的性能,表明这个方法能够适用于二倍体生物,并且应用于多种组织。当前关于鱼类的eQTL研究较少,主要受限于高昂的测序成本。但是通过本方法,可以在较低成本的情况下挖掘出更多的显著性eQTL。因此,相信本研究在未来的水产行业会有更加出色的发挥。

[参 考 文 献]

- [1] LONSDALE J, THOMAS J, SALVATORE M, et al. The genotype-tissue expression (GTE_x) project [J]. *Biopreserv Biobank*, 2013, 45(6): 580-585.
- [2] CONSORTIUM G. The GTE_x Consortium atlas of genetic regulatory effects across human tissues [J]. *Science*, 2020, 369(6509): 1318-1330.
- [3] CASTEL S E, AGUET F, MOHMMADI P, et al. A vast resource of allelic expression data spanning human tissues [J]. *Genome Biol*, 2020, 21(1): 234.
- [4] ZHANG Y, QUICK C, YU K, et al. PTWAS: investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis [J]. *Genome Biol*, 2020, 21(1): 232.
- [5] SHABALIN A A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations [J]. *Bioinformatics*, 2012, 28(10): 1353-1358.
- [6] ONGEN H, BUIL A, BROWN A A, et al. Fast and efficient QTL mapper for thousands of molecular phenotypes [J]. *Bioinformatics*, 2016, 32(10): 1479-1485.
- [7] HUANG H, FANG M, JOSTINS L, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution [J]. *Nature*, 2017, 547(7662): 173-178.
- [8] MOMOZAWA Y, DMITRIEVA J, THEATRE E, et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes [J]. *Nat Commun*, 2018, 9(1): 2427.
- [9] BAHCALL O G. Human genetics: GTE_x pilot quantifies eQTL variation across tissues and individuals [J]. *Nat Rev Genet*, 2015, 16(7): 375.
- [10] TURLEY P, WALTERS RK, MAGHZIAN O, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG [J]. *Nat Genet*, 2018, 50(2): 229-237.
- [11] CHANG C C, CHOW C C, TELLIER L C, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets [J]. *Gigascience*, 2015, 4(1): 7.
- [12] ZHOU X, STEPHENS M. Efficient multivariate linear mixed model algorithms for genome-wide association studies [J]. *Nat Methods*, 2014, 11(4): 407-409.
- [13] ROBERT C P, CHOPIN N, ROUSSEAU J. Harold Jeffroys's theory of probaility revisited [J]. *Statist Sci*, 2009, 24(2): 141-172.
- [14] SUL J H, HAN B, YE C, et al. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches [J]. *PLoS Genet*, 2013, 9(6): e1003491.
- [15] FLUTRE T, WEN X, PRITCHARD J, et al. A statistical framework for joint eQTL analysis in multiple tissues [J]. *PLoS Genet*, 2013, 9(5): e1003486.
- [16] ACHARYA C R, MCCARTHY J M, OWZAR K, et al. Exploiting expression patterns across multiple tissues to map expression quantitative trait loci [J]. *BMC Bioinformatics*, 2016, 17: 257.
- [17] 吴泓杉. eQTL检测的新方法及应用[D]. 厦门: 集美大学, 2022.

(责任编辑 朱雪莲 英文审校 黄力行)