北京林业大学学报
JOURNAL OF BEIJING FORESTRY UNIVERSITY

功能作图的混合效应模型开发

王子健 叶梅霞 张涵 邬荣领

Mixed-effect model development for functional mapping

Wang Zijian, Ye Meixia, Zhang Han, Wu Rongling

引用本文:
王子健, 叶梅霞, 张涵, 邬荣领. 功能作图的混合效应模型开发[J]. 北京林业大学学报, 2024, 46(5):163-172. doi: 10.12171/j.1000-1522.20220416
Wang Zijian, Ye Meixia, Zhang Han, Wu Rongling. Mixed-effect model development for functional mapping[J]. Journal of Beijing Forestry University, 2024, 46(5):163-172. doi: 10.12171/j.1000-1522.20220416

在线阅读 View online: https://doi.org/10.12171/j.1000-1522.20220416

您可能感兴趣的其他文章

Articles you may be interested in

基于Possion回归混合效应模型的长白落叶松一级枝数量模拟

First-order branch number simulation for *Larix olgensis* plantation through Poisson regression mixed effect model

北京林业大学学报, 2017, 39(11): 45-55. https://doi.org/10.13332/j.1000-1522.20170204

基于混合效应的杂种落叶松人工幼龄林单木枯损模型

Individual tree mortality model for hybrid larch young plantations based on mixed effects

北京林业大学学报, 2018, 40(10): 1-10. https://doi.org/10.13332/j.1000-1522.20170437

基于混合效应模型的新疆天山云杉单木胸径预测模型构建

Predicting model construction of single tree DBH of *Picea schrenkiana* in Xinjiang of northwestern China based on mixed effects model

北京林业大学学报, 2020, 42(7): 12-22. https://doi.org/10.12171/j.1000-1522.20190236

基于混合效应模型的人工红松枝下高模型研建

Construction of the height to crown base mixed model for Korean pine

北京林业大学学报, 2020, 42(9): 28-36. https://doi.org/10.12171/j.1000-1522.20190366

基于混合效应的长白落叶松一级枝条基径预估模型

Prediction model of base diameter of primary branch for *Larix olgensis* based on mixed effects

北京林业大学学报, 2020, 42(10): 1-10. https://doi.org/10.12171/j.1000-1522.20200133

胡杨幼苗生长相关性状QTL上位性分析

QTL epistasis effect analysis of seedling growth-related traits in *Populus euphratica*

北京林业大学学报, 2018, 40(12): 49-59. https://doi.org/10.13332/j.1000-1522.20180332

DOI:10.12171/j.1000−1522.20220416

功能作图的混合效应模型开发

王子健　叶梅霞　张　涵　邬荣领
（北京林业大学生物科学与技术学院，计算生物学中心，北京 100083）

摘要:【目的】以大肠杆菌菌株丰度和功能作图模型为研究基础，通过在功能作图模型中引入群体的固定效应和个体间亲缘关系造成的随机效应，探究混合效应对功能作图模型性能的影响。**【方法】**本研究在功能定位框架基础上，使用大肠杆菌动态培养的生长数据作为实际案例，将亚群和 SNP 基因型作为固定效应的来源，将固定效应因素融入定位模型，提出了 $\boldsymbol{Q}$ 矩阵模型的扩展；在保留使用方差–协方差模型对随机残差建模的前提下，使用 Legendre 模型对随机效应建模，开展了固定效应加一般性的随机效应的混合效应模型分析（模型 1）；利用限制性似然估计的方法推导其方差–协方差参数、随机效应和固定效应，开展了固定效应和亲缘关系造成随机效应的混合模型（模型 2）分析；利用 Zwald 检验法推导各标记位点 p 值的计算方法。**【结果】**(1)2 种模型中，95% 的标记呈现出 p 值与期望值相吻合的特点，p 值分布在 QQ 图中的上翘结果满意。(2)模型 2 相比模型 1 检测到了更多的 SNP 位点，表明模型 2 对亲缘关系造成的随机效应解释性更强。(3)计算机模拟结果显示：样本量较小、遗传力较低时，模型假阳性率为 4.77%；当样本量为 800，遗传力为 1% 时，模型对 QTL 的发现率可超过 70%；或当样本量为 400，遗传力超过 1.5% 时，QTL 发现率也可超过 70%。**【结论】**本研究提出的在功能作图模型中引入固定效应和亲缘关系造成的随机效应的混合模型的方法较好地完善了功能定位理论，对固定效应中的协变量因素具有较好的校正功能，可有效剖分随机效应与剩余残差；为后续完善功能定位，开发固定效应加亲缘关系（$\boldsymbol{Q}+\boldsymbol{K}$）模型的软件包奠定了良好基础。

关键词：功能作图；关联分析；混合效应；限制性最大似然估计；数量性状位点
中图分类号：Q943　　**文献标志码：**A　　**文章编号：**1000−1522(2024)05−0163−10
引文格式：王子健，叶梅霞，张涵，等. 功能作图的混合效应模型开发 [J]. 北京林业大学学报，2024，46(5)：163−172. Wang Zijian, Ye Meixia, Zhang Han, et al. Mixed-effect model development for functional mapping[J]. Journal of Beijing Forestry University, 2024, 46(5): 163−172.

Mixed-effect model development for functional mapping

Wang Zijian　Ye Meixia　Zhang Han　Wu Rongling
(Center of Computational Biology, School of Biological Sciences and Biotechnology,
Beijing Forestry University, Beijing 100083, China)

Abstract: [**Objective**] Using the abundance of *Escherichia coli* strains and functional mapping model as the research foundation, this study explored the impact of mixed effects on the performance of functional mapping model by introducing fixed effects of the population and random effects caused by kinship relationship among individuals into the functional mapping model. [**Method**] Based on the framework of functional mapping, this study employed the growth data from dynamic cultures of *Escherichia coli* as a practical case. Subgroups and SNP genotypes were considered as sources of fixed effects, and these fixed effect factors were integrated into the mapping model, leading to the extension of Q-matrix model. While maintaining the use of variance-covariance model for modeling random residuals, the Legendre model was employed to model random effects. A mixed-effect model analysis combining fixed effects with general random effects (model 1) was conducted. Additionally, the restricted maximum likelihood estimation method was utilized to derive variance-covariance parameters, random effects, and fixed effects, enabling

收稿日期: 2022−10−17　**修回日期:** 2023−02−16
基金项目: 国家自然科学基金项目（32071796）。
第一作者: 王子健。主要研究方向：系统生物学。Email: 1614413741@qq.com　地址：100083 北京市海淀区清华东路 35 号北京林业大学。
责任作者: 叶梅霞，博士，副教授。主要研究方向：数量遗传学与计算生物学。Email: yemeixia123@bjfu.edu.cn　地址：同上。
本刊网址: http://j.bjfu.edu.cn; http://journal.bjfu.edu.cn

the analysis of a mixed model combining fixed effects with random effects arising from kinship relationships (model 2). Finally, the Zwald test method was utilized to derive the calculation method for *p*-values at each marker locus. [**Result**] (1) In both models, 95% of the markers exhibited *p*-values that were consistent with the expected values, resulting in satisfactory upward curvature in the QQ plot. (2) Compared with model 1, model 2 detected a greater number of SNP loci, indicating that model 2 provided a stronger explanation for the random effects caused by kinship relationship. (3) Computer simulation results revealed that when the sample size was small and the heritability was low, the false-positive rate of the model was 4.77%. However, when the sample size reached 800 and the heritability was 1%, the discovery rate of quantitative trait loci (QTL) by the model can exceed 70%. Alternatively, when the sample size was 400 and the heritability exceeded 1.5%, the QTL discovery rate can also exceed 70%. [**Conclusion**] The mixed model approach proposed in this study, which introduces fixed effects and random effects caused by kinship relationship into the functional mapping model, effectively improves the theory of functional localization. This approach exhibits excellent calibration capabilities for covariate factors in fixed effects and can effectively dissect random effects from remaining residuals. This lays a solid foundation for subsequent improvements in functional localization and the development of software packages for the fixed effects plus kinship (***Q*** + ***K***) model.

Key words: functional mapping; association analysis; mixed effect; restricted maximum likelihood estimation; quantitative trait loci (QTL)

动态性状是一类随生长时间点变化，其表型数值不断改变的性状，比如动植物在其生长和发育过程中具有明显的时间性，鸟类的产蛋率、哺乳动物的产奶量、拟南芥的株高和茎长等多属于动态性状[1−2]。动态性状在帮助人们深刻理解复杂性状的形成机理方面起到了重要作用[2]。Ma 等[3]提出了功能作图的分析方法，使用了与时间相关的生物生长函数来进行分析，获得了随时间变化的与生物生长相关的数量性状位点（quantitative trait loci，QTL）。研究动态性状有助于揭示基因的动态表达，为数量性状遗传研究提供理论依据。通过对动态性状的分析计算，研究者们可以获得基因对性状的调控作用、调控作用的动态变化、调控效应的变化、遗传控制持续的时间等方面的信息。因而，利用动态性状开展 QTL 定位将对复杂性状的遗传结构解析发挥重要作用[4]。

作为下一代 QTL 定位统计工具，功能作图对解析以上问题具有较好的统计功效。与传统基于静态单时间点表型的 QTL 定位技术相比，功能作图采用生物学生长曲线来拟合动态性状的生长过程，同时采用方差–协方差模型对不同时间点之间的性状相关性进行建模，其合理考虑了生物学发育原理，对解析 QTL 如何控制复杂性状之谜发挥了重要作用[5]。根据动态作用模式，功能作图总结了 4 类发育 QTL：早–晚型、快–慢型、短–长型、连续型，即不同基因型的个体在发育过程中呈现出显著的差异，这些差异体现在个体发育起始时间的早晚、发育速率的快慢、发育调控时间的长短，以及个体在生长前期和生长后期的发育速率恰好相反等方面。该技术自 2002 年诞生后[3]就一直受到关注，2 年后就被 Nature Review Genetics 刊物重点推荐[6]，4 年后就以观点论文的形式重新在该刊物上做了介绍[7]。随着表型组时代的来临，利用高通量表型测定设备来获取动态发育表型数据越来越受关注，2015 年有学者着重介绍并分析了功能定位方法在表型组时代的应用前景[8−9]，进一步证明功能定位仍是当前复杂性状解析的主流技术。

时间序列数据是生物动态数据中最为常见的一类，连续时间点之间的表型数据存在较强的自回归性质。针对该问题，功能作图首创性地使用方差–协方差模型，如自回归模型、结构前相依模型，来对多时间点间的协方差开展结构化建模，以此提高了其在模型稳定性、统计功效等方面的性能[10−12]。随后，Hou 等[13]在功能作图基础上，提出使用混合非线性模型（nonlinear mixed-effect model，NLME）的方法，将随机效应与方差–协方差相区别，使功能作图在技术上得到了完善。该方法重点考虑了随机效应、时间序列相关性、测量误差 3 方面效应对不同个体之间呈现的数据变异性的贡献，因而可以准确剖分出 QTL 基因型所发挥的固定效应与随机效应。进一步根据一阶泰勒展开的数学原理，Zhu 等[14]对功能作图技术中的非线性生长曲线作了线性变换，利用了线性混合模型（linear mixed model，LMM）的计算优势，有效降低了混合效应模型在功能作图中的应用难度。

Hou 等[13]的工作，为功能作图的有效扩展提供了线索，推动了功能作图的优化，但是在 QTL 基因

型的曲线拟合问题上，Hou 等仍然使用最大似然估计的统计框架。最大似然估计的统计框架虽然具有相合性、同变性、渐近正态性等性质，但是并不能无偏估计混合非线性模型中的方差分量，因而影响生长曲线参数的准确估计[15]。此外，在全基因组关联分析（genome-wide association studies，GWAS）研究中，结构化群体的固定效应和亲缘关系带来的随机效应会影响分析结果[16]。首先，结构化群体的固定效应，如亚群差异、环境因素、年龄、性别等都可能对表型产生显著影响。在 GWAS 分析中，如果未能妥善考虑这些固定效应，会造成分析结果的不准确。其次，亲缘关系造成的随机效应在 GWAS 研究中也是一个不可忽略的因素。群体间个体存在着复杂的亲缘关系，使得群体中的个体与个体之间并非独立存在，进而在关联分析中产生潜在的偏差，这种由亲缘关系导致的随机效应会导致假阳性的结果，进而影响 GWAS 分析的可信度和准确度。Hou 等的方法未涉及到固定效应和亲缘关系造成的随机效应，因而适用群体类型较为有限，比如人工杂交群体、重组自交系群体等。从生长曲线参数角度而言，其曲线形状受一系列潜在的协变量影响，如年龄、性别、种族等因素，在功能作图统计框架中如何在固定效应的基础上引入合理的协变量，也将成为动态性状 QTL 定位技术的关键环节。

综合这 3 个问题，即最大似然估计对方差分量的有偏估计、功能定位的固定效应、亲缘关系造成的随机效应，我们认为，使用限制性最大似然估计（restricted maximum likelihood，REML）对方差分量、固定效应中的协变量参数、随机效应参数的无偏估计将是利用混合线性模型开展 QTL 定位的重要命题[17]，也是功能定位接下来发展的重要方向。本文在 Hou 等[13]的方法基础上，提出 $\boldsymbol{Q}$ 矩阵的构建方法，对影响动态性状的重要协变量作出模型扩展；同时利用 REML 法对功能作图的混合效应模型开展参数推导，进一步构建 Zwald 统计量[18]，计算 p 值，对是否存在的假设作出检验，为进一步完善功能作图作出贡献。此外，以 100 株大肠杆菌的生长数据作为实例分析，验证所扩展的模型，并以此展示模型功效。

1　框架与方法

1.1　功能作图的混合模型

假设作图群体由 n 个个体组成，不同位点上假定有 J 个 QTL 基因型，设 $\boldsymbol{t}_i=\{\boldsymbol{t}_{i\tau}\}_{\tau=1}^{T}$ 是个体 i 在 T 个时间点测量的时间向量，$\boldsymbol{y}_i=\{\boldsymbol{y}_i(\boldsymbol{t}_{i\tau})\}_{\tau=1}^{T}$ 是个体 i 的纵向表型测量向量。以逻辑斯谛生长曲线为例，个体 i 的表型性状可表示为 $\boldsymbol{y}_i=\sum_{j=1}^{J}\varepsilon_{ij}g\left(\boldsymbol{A}_i,\boldsymbol{B}_j;\boldsymbol{t}_i\right)+\boldsymbol{s}_i+\boldsymbol{e}_i$，其中 ε_{ij} 是指示变量，如果个体 i 携带 QTL 基因型 j，则定义为 1，否则为 0；g 是 $\boldsymbol{A}_i$、$\boldsymbol{B}_j$ 和 $\boldsymbol{t}_i$ 的非线性函数，其中 $\boldsymbol{A}_i$ 是个体 i 的协变量关于曲线效应的参数向量，$\boldsymbol{B}_j$ 代表了基因型 j 关于生长曲线的参数向量。$\boldsymbol{s}_i$ 是个体 i 的随机效应向量，因随机效应不具有固定的曲线形式，因此采用 Legendre orthogonal polynomials（LOP）曲线[19]对其建模。LOP 是一系列基函数与各阶系数的线性组合，以 $\boldsymbol{\varphi}(\boldsymbol{t}_i)$ 代表各阶基函数，u_k 是各阶基函数的系数，则 $\boldsymbol{s}_i=\sum_{k=1}^{K}\boldsymbol{\varphi}_k(\boldsymbol{t}_i)\cdot u_k$。$\boldsymbol{e}_i$ 是一个 $(1\times T)$ 的误差项向量，通常服从正态分布，均值为零向量。个体的方差–协方差矩阵为 $\boldsymbol{\Sigma}_{\boldsymbol{e}}$，是一个 $(T\times T)$ 的矩阵，可由一组参数对其建模，代表个体在不同时间点上的表型自相关关系。

1.2　功能作图的 $\boldsymbol{Q}$ 模型扩展

对曲线参数作固定效应的来源分解，具体包括：均值背景参数向量、群体结构如亚群效应部分的参数向量、性别或年龄等其他协变量的效应参数向量、加性或显性效应的曲线参数向量。具体来说：用 $\boldsymbol{\beta}_{\text{fix}}$ 来表示参数向量类型 1；若以 $\boldsymbol{Q}$ 矩阵表示群体的亚群信息，$\boldsymbol{\beta}_{\boldsymbol{Q}_1}\cdots\boldsymbol{\beta}_{\boldsymbol{Q}_n}$ 表示所有亚群的曲线参数向量，即类型 2；对于类型 3，如果以年龄、性别为例构成协变量，以 $\boldsymbol{\beta}_{\text{age}}$ 和 $\boldsymbol{\beta}_{\text{sex}}$ 表示年龄和性别协变量的曲线参数向量；对于类型 4，分别用 $\boldsymbol{\beta}_{\text{a}}$ 和 $\boldsymbol{\beta}_{\text{d}}$ 来表示生长曲线中的加性效应和显性效应参数向量。其中类型 1 的参数可以合并至类型 2 中，$\boldsymbol{\beta}$ 表示参数向量，以逻辑斯谛曲线为例，$\boldsymbol{\beta}_{\text{age}}=(a_{\text{age}},b_{\text{age}},r_{\text{age}})$，$\boldsymbol{\beta}_{\text{a}}=(a_{\text{a}},b_{\text{a}},r_{\text{a}})$，依此类推，各因子的 $\boldsymbol{\beta}$ 向量所包含的元素个数与曲线参数个数相等。因此：

$$\boldsymbol{\beta}=((\boldsymbol{\beta}_{\boldsymbol{Q}_1}\cdots\boldsymbol{\beta}_{\boldsymbol{Q}_n},\boldsymbol{\beta}_{\text{age}},\boldsymbol{\beta}_{\text{sex}}),(\boldsymbol{\beta}_{\text{a}},\boldsymbol{\beta}_{\text{d}})) \tag{1}$$

同样地，建立

$$\boldsymbol{Q}_i=(\boldsymbol{Q}_{1i}\cdots\boldsymbol{Q}_{ni},\boldsymbol{E}_{\text{age}i},\boldsymbol{F}_{\text{sex}i},\boldsymbol{M}_{\text{add}i},\boldsymbol{N}_{\text{dom}i}) \tag{2}$$

来表示每个个体的所有固定因素向量，包含各协变量在内。其中 $\boldsymbol{Q}_{1i}\cdots\boldsymbol{Q}_{ni}$ 表示个体被分配到的概率为 $q_1\cdots q_n$。$\boldsymbol{E}_{\text{age}}$ 是年龄向量，可处理为分类变量；$\boldsymbol{F}_{\text{sex}},\boldsymbol{M}_{\text{add}},\boldsymbol{N}_{\text{dom}}$ 是指示变量。若个体 i 的基因型为 aa，$\boldsymbol{M}_{\text{add}i}=0$，$\boldsymbol{N}_{\text{dom}i}=0$；如果基因型为 Aa，$\boldsymbol{M}_{\text{add}i}=0$，$\boldsymbol{N}_{\text{dom}i}=1$；如果基因型为 AA，$\boldsymbol{M}_{\text{add}i}=1$，$\boldsymbol{N}_{\text{dom}i}=0$。因此，$\boldsymbol{y}_i$ 的性状构成可由以上 $\boldsymbol{Q}$ 模型改写为

$$\boldsymbol{y}_i=g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_i)+\sum_{k=1}^{K}\boldsymbol{\varphi}_k(\boldsymbol{t}_i)u_k+\boldsymbol{e}_i \tag{3}$$

1.3　非线性曲线的线性化展开

为将方程（3）线性化，采用一阶泰勒扩展 $g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_i)$，$\sum_{k=1}^{K}\boldsymbol{\varphi}_k(\boldsymbol{t}_i)u_k$ 的一阶泰勒展开式与原式相同，方程（3）被线性化为

$$\tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i - \left[g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_i) + \sum_{k=1}^{K}\boldsymbol{\varphi}_k(\boldsymbol{t}_i)u_k\right] + \boldsymbol{W}_i\boldsymbol{\beta} + \sum_{k=1}^{K}\boldsymbol{\varphi}_k(\boldsymbol{t}_i)u_k = \boldsymbol{W}_i\boldsymbol{\beta} + \sum_{k=1}^{K}\boldsymbol{\varphi}_k(\boldsymbol{t}_i)u_k + \boldsymbol{e}_i \tag{4}$$

$\boldsymbol{W}_i$由随时间变化的元素组成，

$$\boldsymbol{W}_i(\boldsymbol{t}_i) = \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_i)}{\delta(\boldsymbol{\beta}^{\mathrm{T}})} \tag{5}$$

其中$\boldsymbol{W}_i(\boldsymbol{t}_i)$是个体 i 关于生长曲线各参数的一阶导数矩阵，是一个$(T\times I)$矩阵，其中 $p=\left(n_{\boldsymbol{Q}}+2\right)n_c$，$T$ 是时间点的个数，$n_{\boldsymbol{Q}}$表示 $\boldsymbol{Q}$ 矩阵的列数，n_c表示曲线的参数个数。综合影响固定效应的所有因子，$\boldsymbol{W}_i(\boldsymbol{t}_i)$可以表示为

$$\boldsymbol{W}_i(\boldsymbol{t}_i) = \left(\boldsymbol{W}_{i(\boldsymbol{Q}_1)}(\boldsymbol{t}_i),\cdots,\boldsymbol{W}_{i(\boldsymbol{Q}_n)}(\boldsymbol{t}_i),\boldsymbol{W}_{i(\boldsymbol{E}_{\mathrm{age}})}(\boldsymbol{t}_i),\boldsymbol{W}_{i(\boldsymbol{F}_{\mathrm{sex}})}(\boldsymbol{t}_i),\boldsymbol{W}_{i(\boldsymbol{M}_{\mathrm{add}})}(\boldsymbol{t}_i),\boldsymbol{W}_{i(\boldsymbol{N}_{\mathrm{dom}})}(\boldsymbol{t}_i)\right) =$$

$$\begin{pmatrix}
\frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i1})}{\delta(\boldsymbol{\beta}_{\boldsymbol{Q}_1})^{\mathrm{T}}} & \cdots & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i1})}{\delta(\boldsymbol{\beta}_{\boldsymbol{Q}_n})^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i1})}{\delta(\boldsymbol{\beta}_{\mathrm{age}})^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i1})}{\delta(\boldsymbol{\beta}_{\mathrm{sex}})^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i1})}{\delta(\boldsymbol{\beta}_a)^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i1})}{\delta(\boldsymbol{\beta}_d)^{\mathrm{T}}} \\
\frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i2})}{\delta(\boldsymbol{\beta}_{\boldsymbol{Q}_1})^{\mathrm{T}}} & \cdots & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i2})}{\delta(\boldsymbol{\beta}_{\boldsymbol{Q}_n})^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i2})}{\delta(\boldsymbol{\beta}_{\mathrm{age}})^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i2})}{\delta(\boldsymbol{\beta}_{\mathrm{sex}})^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i2})}{\delta(\boldsymbol{\beta}_a)^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{i2})}{\delta(\boldsymbol{\beta}_d)^{\mathrm{T}}} \\
\vdots & & \vdots & \vdots & \vdots & \vdots & \vdots \\
\frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{iT})}{\delta(\boldsymbol{\beta}_{\boldsymbol{Q}_1})^{\mathrm{T}}} & \cdots & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{iT})}{\delta(\boldsymbol{\beta}_{\boldsymbol{Q}_n})^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{iT})}{\delta(\boldsymbol{\beta}_{\mathrm{age}})^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{iT})}{\delta(\boldsymbol{\beta}_{\mathrm{sex}})^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{iT})}{\delta(\boldsymbol{\beta}_a)^{\mathrm{T}}} & \frac{\delta g(\boldsymbol{Q}_i\boldsymbol{\beta};\boldsymbol{t}_{iT})}{\delta(\boldsymbol{\beta}_d)^{\mathrm{T}}}
\end{pmatrix} \tag{6}$$

令 $K=4$，即采用 4 阶 LOP 模型拟合随机效应，同时令$\boldsymbol{Z}_i=(\boldsymbol{\varphi}_0(t_i),\ \boldsymbol{\varphi}_1(t_i),\cdots,\boldsymbol{\varphi}_4(t_i))$，其中 $\boldsymbol{\varphi}_0(t_i)=1$，$\boldsymbol{\varphi}_1(t_i)=t_i$，$\boldsymbol{\varphi}_2(t_i)=0.5(3t_i^2-1)$，$\boldsymbol{\varphi}_3(t_i)=0.5(5t_i^3-3t_i)$，$\boldsymbol{\varphi}_4(t_i)=0.125(35t_i^3-30t_i^2+3)$。

整合所有个体的线性化模型，表示为

$$\begin{bmatrix}\tilde{\boldsymbol{y}}_1\\ \tilde{\boldsymbol{y}}_2\\ \vdots\\ \tilde{\boldsymbol{y}}_n\end{bmatrix} = \begin{bmatrix}\boldsymbol{W}_1\\ \boldsymbol{W}_2\\ \vdots\\ \boldsymbol{W}_n\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{Z}_1\\ \boldsymbol{Z}_2\\ \vdots\\ \boldsymbol{Z}_n\end{bmatrix}\boldsymbol{u} + \begin{bmatrix}\boldsymbol{e}_1\\ \boldsymbol{e}_2\\ \vdots\\ \boldsymbol{e}_n\end{bmatrix} \tag{7}$$

进一步将矩阵简化为

$$\tilde{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \tag{8}$$

1.4 方差–协方差建模

在限制性最大似然框架下，可以使用 SAD 模型对方差–协方差进行建模[20]。基于 SAD 模型构建，协方差矩阵$\boldsymbol{\Sigma}_{\boldsymbol{e}}$可以表示为

$$\sigma_i^2(\boldsymbol{t}) = \frac{1-\boldsymbol{\varphi}_i^{2\boldsymbol{t}}}{1-\boldsymbol{\varphi}_i^2}v_i^2 \tag{9}$$

$$\sigma_i(t_1,t_2) = \boldsymbol{\phi}_i^{t_2-t_1}\frac{1-\boldsymbol{\phi}_i^{2t}}{1-\boldsymbol{\phi}_i^2}v_i^2 \tag{10}$$

式中：$\sigma_i^2(\boldsymbol{t})$表示性状 i 在时间 t 的方差，$\sigma_i(t_1,t_2)$表示性状 i 在时间$\boldsymbol{t}_1$和$\boldsymbol{t}_2$的协方差，通过设置参数$\boldsymbol{v}_i^2$可以反映测量的误差波动，$\boldsymbol{\phi}$为相关系数向量。

1.5 最大似然估计与限制性最大似然估计

1.5.1 目标对数似然函数

若给定未经线性化的非线性生长曲线，目标对数似然函数可以表述为

$$L(\boldsymbol{\Sigma}_{\boldsymbol{e}}) = \sum_{i=1}^{n}\ln(f(\boldsymbol{y}_i)) \tag{11}$$

而在给定线性化公式(8)的前提下，限制性对数似然函数可以表示为

$$L(\boldsymbol{\Sigma}_{\boldsymbol{e}}) = -\frac{nT}{2}\ln(2\pi) + \frac{1}{2}\ln\left|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\right| - \frac{1}{2}\ln|\boldsymbol{H}| - \frac{1}{2}\ln\left|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{W}\right| - \frac{1}{2}\ln\left|\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{y}\right| \tag{12}$$

$$\boldsymbol{H} = \mathrm{var}(\tilde{\boldsymbol{y}}) = \boldsymbol{Z}\mathrm{var}(\boldsymbol{u})\boldsymbol{Z}^{\mathrm{T}} + \boldsymbol{I}_n\otimes\boldsymbol{\Sigma}_{\boldsymbol{e}} \tag{13}$$

$$\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{W}\left[\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{W}\right]^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}^{-1} \tag{14}$$

式中：$\boldsymbol{H}$ 为群体的协方差矩阵，$\boldsymbol{P}$ 为限制性对数似然函数中间计算过程。

1.5.2 Zwald 检验

Zwald 检验是一套检验 SNP 所属的$\boldsymbol{\beta}$向量是否显著区别于 0 的检验方法。从式(13)中，可以看出 $\boldsymbol{H}$ 是一个由 n 个大小为$(T\times T)$的矩阵填充的对角线块矩阵，则：

$$|\boldsymbol{H}| = \sum_{l=1}^{n}|\boldsymbol{V}_l| = \sum_{l=1}^{n}\left|\boldsymbol{Z}_i\mathrm{var}(\boldsymbol{u})\boldsymbol{Z}_i^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{e}}\right| \tag{15}$$

$$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{W} = \sum_{l=1}^{n}\left[\boldsymbol{W}_l^{\mathrm{T}}\boldsymbol{V}_l^{-1}\boldsymbol{W}_l\right] \tag{16}$$

$$\boldsymbol{H}^{-1} = \mathrm{diag}(\boldsymbol{V}_1^{-1},\boldsymbol{V}_2^{-1},\cdots,\boldsymbol{V}_n^{-1}) \tag{17}$$

$\boldsymbol{\Sigma}_{\boldsymbol{e}}$中的参数$\phi$和$v$可以根据目标函数(12)，采用 Nelder-Mead 的数值解算法[21]或 Newton-Raphson 算法[22]来估计。对于当前的θ估计值，根据 Laird 和 Ware 等推导的参数估计算法[23]，$\boldsymbol{\beta}$可估计为

$$\boldsymbol{\beta}(\theta)=\left[\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{W}\right]^{-1}\cdot\left[\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{W}\right] \tag{18}$$

$$\mathrm{var}(\boldsymbol{\beta})=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{W} \tag{19}$$

根据公式 16～17 的结果，挑选其中关于 SNP 参数的部分数值和亚矩阵，其 Z 值统计量的计算方法为

$$Z=(\boldsymbol{\beta}_{\mathrm{SNP}})^{\mathrm{T}}(\mathrm{var}(\boldsymbol{\beta}_{\mathrm{SNP}}))^{-1}\boldsymbol{\beta}_{\mathrm{SNP}} \tag{20}$$

同样，Z 统计量通过卡方检验转化为 p 值。

1.5.3　随机效应的模型建立

利用上述公式，本文拟开展 2 类随机效应的计算，除固定效应外，第 1 种是考虑个体间随机性而不考虑亲缘关系的随机效应，即一般性的随机效应，见上述公式即可；第 2 种是由亲缘关系造成的随机效应，即 $\boldsymbol{Q}+\boldsymbol{K}$ 模型。第 2 种需要利用分子标记数据建立亲缘关系矩阵 $\boldsymbol{K}$[24]，如：

$$\boldsymbol{K}=\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}}/2\sum p_i(1-p_i) \tag{21}$$

p_i 是第 i 个位点等位基因 a 的频率，$\boldsymbol{S}$ 是标记矩阵，$\boldsymbol{S}$ 的计算方法为

$$\boldsymbol{S}_i=\begin{cases}2-2p_i,\left(\text{若个体}i\text{基因型为aa}\right)\\1-2p_i,\left(\text{若个体}i\text{基因型为Aa}\right)\\0-2p_i,\left(\text{若个体}i\text{基因型为AA}\right)\end{cases} \tag{22}$$

在将 $\boldsymbol{K}$ 引入公式(12)时，需要对其作特征分解，即 $\boldsymbol{K}=\boldsymbol{U}_k\boldsymbol{D}_k\boldsymbol{U}_k^{\mathrm{T}}$，其中 $\boldsymbol{U}_k$ 是特征向量构成的矩阵，$\boldsymbol{D}_k$ 是由特征值构成的对角线矩阵，并且 $\boldsymbol{U}_k$ 具有 $\boldsymbol{U}_k^{-1}=\boldsymbol{U}_k^{\mathrm{T}}$ 的特性。利用该性质，在公式(8)中，使用 $\boldsymbol{U}_t\boldsymbol{W}=(\boldsymbol{U}_k\otimes\boldsymbol{I}_T)\boldsymbol{W}$ 代替 $\boldsymbol{W}$，使用 $\boldsymbol{U}_t\boldsymbol{y}=(\boldsymbol{U}_k\otimes\boldsymbol{I}_T)\boldsymbol{y}$ 代替 $\boldsymbol{y}$，使用 $\boldsymbol{U}_t\boldsymbol{Z}=(\boldsymbol{U}_k\otimes\boldsymbol{I}_T)\boldsymbol{Z}$ 代替 $\boldsymbol{Z}$。

1.6　功能注释

在本研究中，为了进一步了解显著位点参与的生物学过程，从大肠杆菌基因组注释文件中获取每个 SNP 位点所在的基因编号，对 p 值小于阈值线的位点，使用 NCBI 数据库查询其生物学含义进行功能注释。

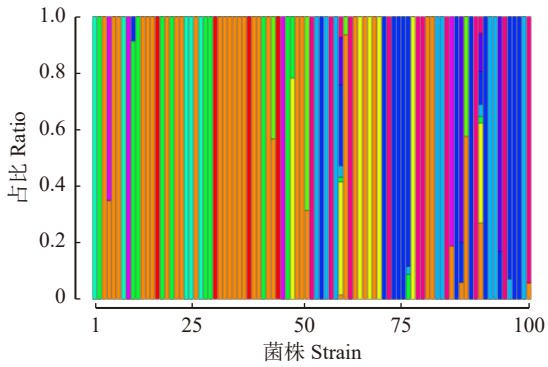
2　应用与分析

2.1　数据概况

本研究采用 100 个大肠杆菌（*Escherichia coli*）菌株在 36 h 内的动态培养数据对上述模型展开实例分析。所使用的菌株数据来自 Yang 等[25]，其中大肠杆菌菌株收集自国家微生物资源平台。本研究使用的大肠杆菌动态生长数据来自单独培养条件下的各菌株，各菌株分别设置 3 次独立重复试验，培养条件参见 Jiang[26]等。单独培养周期设定为 36 h，于培养的前 2 h 内、2～12 h 内、12～36 h 期间，分别每间隔 0.5 h、2 h、4 h 使用定量 PCR（quantitative PCR，qPCR）技术测定一次菌株丰度，共获得 14 次动态生长数据。使用 Illumina Hiseq 测序平台对 100 个大肠杆菌菌株开展全基因组测序，由北京奥维森基因公司完成，将测序数据与参考基因组进行比对，获得全基因组 SNP 标记。经筛选错误信息、高缺失率标记后，删除缺失基因型数据比率大于 20% 个体的标记，共获得 58 397 个高质量 SNP。

2.2　曲线选择与 $\boldsymbol{Q}$ 矩阵的建立

使用 fastSTRUCTURE 软件对 SNP 数据开展亚群分析[27]，结果揭示出 100 个大肠杆菌菌株可划分为 11 个亚群，并得到各菌株分别归属于 11 个亚群的概率值，利用概率值绘制如下群体结构分析图（图 1）。图中每一列代表一个大肠杆菌菌株，每种颜色的占比表明该菌株划分为各个亚群的概率，累加和为 1。考虑 SNP 基因型由 0 或 1 组成，因而其 $\boldsymbol{Q}$ 矩阵可直接构建为 100 行 × 12 列，每一行代表一个菌株，前 11 列代表各菌株分别归属于各亚群的概率，第 12 列表示 SNP 基因型。根据大肠杆菌在不同时间点的动态变化（图 2），选择逻辑斯谛生长曲线开展模型分析。



每一列代表一个大肠杆菌菌株，每种颜色的占比表示该菌株划分为各个亚群的概率。Each column represents a strain of *Escherichia coli*, and the proportion of each color indicates the probability of the strain being divided into various subgroups.

图 1　100 株大肠杆菌的群体结构分析

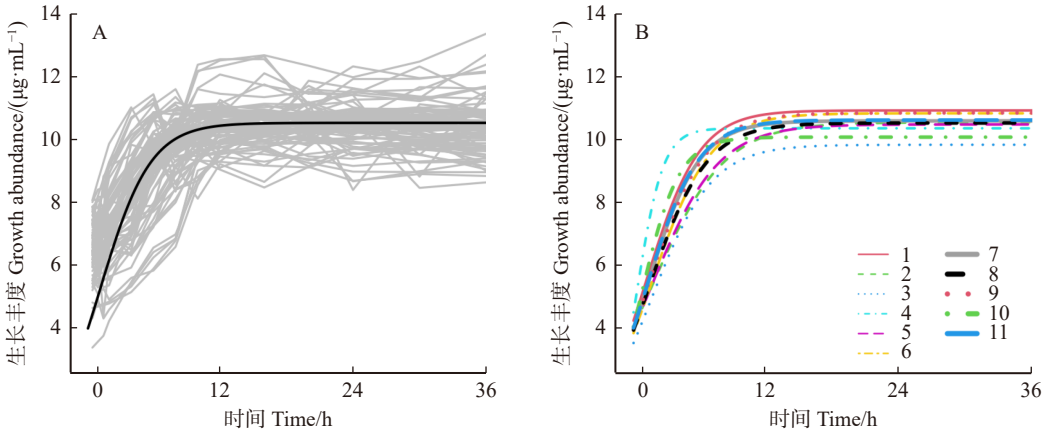
Fig. 1　Analysis of population structure of 100 strains of *Escherichia coli*

2.3　随机效应的模型建立

利用上述公式，本文拟开展 2 类随机效应的计算，除固定效应外，第 1 种是一般性的随机效应，第 2 种是由亲缘关系造成的随机效应，即 $\boldsymbol{Q}+\boldsymbol{K}$ 模型。

2.4　计算机模拟策略

为评价模型性能，进一步通过设定不同样本量、不同遗传力，对本文模型进行检验。在大肠杆菌实际数据基础上，继续开展基于实际数据的计算机模拟。模拟分为如下步骤：开展单时间点的 QTL 分析，寻找非关联的 SNP 位点，即不显著位点；综合单时间点和整体动态的 QTL 分析结果，选出在 2 类 QTL



A. 大肠杆菌菌株在 0～36 h 的丰度动态变化曲线；B. 不同亚群之间的菌株平均丰度差异。A, dynamic change curves of abundance of *E. coli* strains in 0−36 h; B, difference in average abundance of strains among different sub-populations.

图 2 100 个大肠杆菌菌株的生长动态

Fig. 2 Growth dynamics of 100 stains of *Escherichia coli*

定位中都不显著的 SNP 位点；对步骤 2 选出的 SNP 按照等基因组位置间隔的原则，从中挑选 100 个位点，以最大程度上减少相邻 SNP 间的连锁不平衡现象；在原数据中通过有放回抽样的方法设置 100、200、400、600、800 共 5 种不同样本量，设置遗传力 0.0%、0.25%、0.5%、1.0%、1.5%、2.0% 共 6 种情况，不同样本量和不同遗传力的组合共 30 组；针对每一种样本量和遗传力的组合，模拟 SNP 效应，将所模拟的 SNP 动态效应重新整合至原个体的动态表型，随后再利用不考虑随机效应和采用第 2 类随机效应模型的算法分别开展实际计算。当遗传力为 0.0% 时，统计显著 SNP 出现的比率，记为模型假阳性率；当遗传力大于 0.0% 时，统计显著 SNP 出现的比率，记为模型功效。

3 结果与分析

3.1 生长动态分析

基于 100 个大肠杆菌菌株 36 h 内的生长动态（图 2A），使用逻辑斯谛曲线，利用最大似然法对 100 个菌株的丰度动态进行生长拟合。利用曲线参数，同时结合 Jiang 等[28]对逻辑斯谛曲线的异时性解析，计算出该菌株群体在 5.56 h 出现曲线的生长拐点。考虑该群体具有较强的分层现象，为了检验不同亚群间的菌株丰度动态是否存在统计学显著差异，分别计算各亚群逻辑斯谛生长曲线的 a、b、r 参数和不考虑亚群效应的逻辑斯谛生长曲线的 a、b、r 参数（表 1），并进一步开展逐个参数的单因素方差分析。结果表明：a、b、r 参数在不同亚群之间均存在显著差异，p 值分别为 0.000 72，0.002 30，0.000 58，显著低于 0.01。图 2B 显示：在渐近生长量角度，亚群 1 的菌株丰度最大值达到 10.95 ± 0.03，处于所有亚群的最高水平；亚群 3 的最大丰度为 9.86 ± 0.02，处于最低水平；在增长速率方面，亚群 4 的增长速率为 0.75，表现最快；亚群 5 的增长速率为 0.22，表现最慢。不同亚群的菌株在生长丰度、生长速率方面呈现出差异，表明亚群效应会影

表 1 不同亚群的逻辑斯谛生长曲线参数

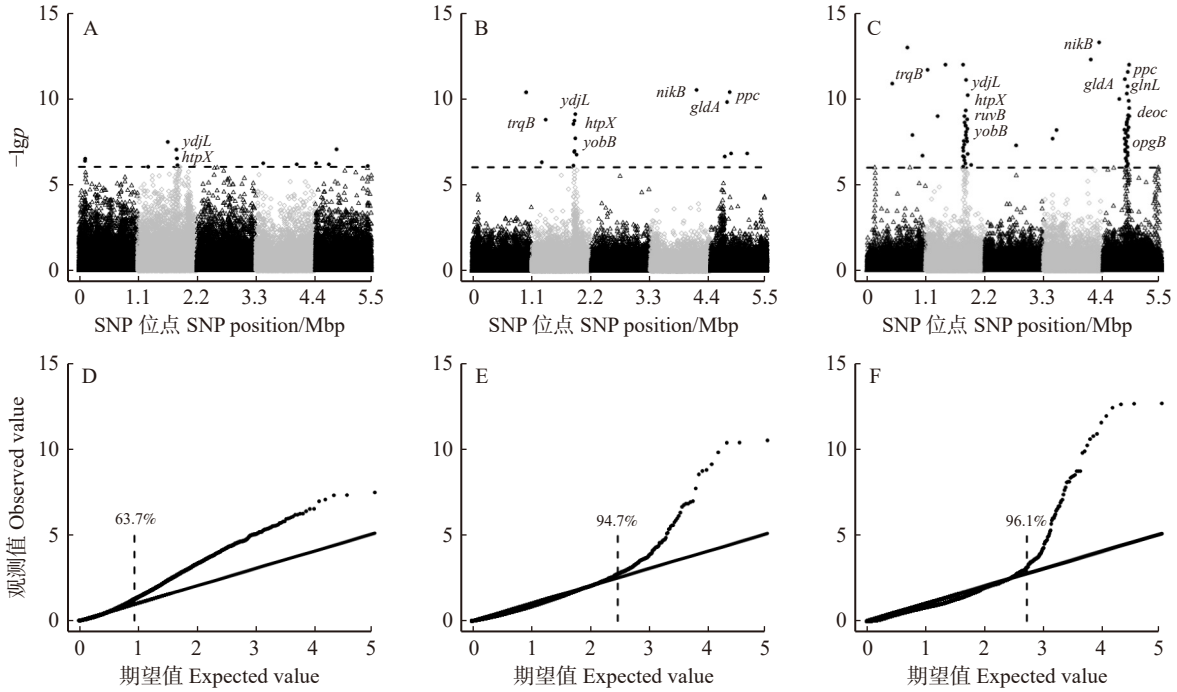
Tab. 1 Parameters of Logistic growth curves for different subgroups

项目 Item		a	b	r
不考虑亚群效应 Not considering subgroup effect		11.030 4	1.871 5	0.569 3
考虑亚群效应 Considering subgroup effect	亚群1 Subgroup 1	10.956 8	1.643 1	0.312 7
	亚群2 Subgroup 2	10.577 0	1.662 6	0.204 1
	亚群3 Subgroup 3	9.862 7	1.853 2	0.265 5
	亚群4 Subgroup 4	10.387 2	1.414 5	0.746 8
	亚群5 Subgroup 5	10.510 3	1.688 8	0.227 6
	亚群6 Subgroup 6	10.864 5	1.887 9	0.274 3
	亚群7 Subgroup 7	10.623 0	1.735 5	0.325 6
	亚群8 Subgroup 8	10.555 8	1.736 1	0.282 3
	亚群9 Subgroup 9	10.876 2	1.751 7	0.283 4
	亚群10 Subgroup 10	10.105 6	1.578 4	0.483 0
	亚群11 Subgroup 11	10.649 8	1.720 6	0.326 8

响大肠杆菌的发育。因此可以认为，除 SNP 基因型外，亚群应被视为协变量，在功能作图模型中应该被处理成固定效应，对曲线参数 a、b、r 均产生影响。在分析过程中，应将这些差异充分融入进 $\boldsymbol{Q}$ 矩阵。

3.2　功能作图模型

功能作图模型不考虑亲缘关系导致的随机效应。图 3A 和图 3D 表示通过功能作图模型分析大肠杆菌动态生长丰度得到的曼哈顿图和 QQ 图。曼哈顿图表明在基因组 1.9 Mbp 附近检测到了一个峰值，然而 QQ 图较为上翘，其中 63.7% 的 SNP 位点的 p 值的观测分布与 p 值的期望分布一致。结合曼哈顿图和 QQ 图分析，忽视了随机效应的功能作图模型表现一般，模型出现了较高的假阳性率。



A、D. 功能作图得到的曼哈顿图和 QQ 图；B、E. 一般性随机效应得到的曼哈顿图和 QQ 图；C、F. $\boldsymbol{Q}+\boldsymbol{K}$ 模型得到的曼哈顿图和 QQ 图。A and D, the Manhattan and QQ plot obtained from functional mapping; B and E, the Manhattan and QQ plot obtained from the general random effect model; C and F, the Manhattan and QQ plot obtained by the $\boldsymbol{Q}+\boldsymbol{K}$ model .

图 3　功能作图模型和 2 种混合效应模型的曼哈顿图和 QQ 图

Fig. 3　Manhattan and QQ plots of functional mapping model and the two mixed-effect models

3.3　一般随机效应模型

一般随机效应是指群体随机效应不考虑由群体亲缘关系造成，属于一般化的随机效应。图 3B 和图 3E 表示通过一般随机效应模型分析大肠杆菌动态生长丰度得到的曼哈顿图和 QQ 图。曼哈顿图揭示在基因组 2.0 Mbp 和 4.7 Mbp 附近检测到了 2 个明显的峰值，峰值最高点为 11.2。联合曼哈顿图与 QQ 图发现：94.7% 的 SNP 位点的 p 值的观测分布与 p 值的期望分布一致，当 SNP 位点的 p 值小于 10^{-2}时，开始尾巴上翘，上翘结果理想。与功能作图模型相比，一般随机效应模型效果显著提升。

3.4　亲缘关系随机效应模型

亲缘关系随机效应模型是指群体内随机效应并不是随机的，而是由个体间的亲缘关系决定造成，此模型又称为 $\boldsymbol{Q}+\boldsymbol{K}$ 模型，即混合效应功能作图模型。本文利用全基因组 SNP 数据构建亲缘关系矩阵，并将其融入模型。图 3C 和图 3F 显示了利用 $\boldsymbol{Q}+\boldsymbol{K}$ 模型得到的分析结果。与一般随机效应模型相比，$\boldsymbol{Q}+\boldsymbol{K}$ 模型得到的曼哈顿图和 QQ 图总体趋势一致。曼哈顿图揭示，在 2.0 Mbp 和 4.7 Mbp 附近检测到了同样的 2 个峰值，但是相比一般随机效应模型，该模型产生的峰值更为明显和突出。结合 QQ 图发现，96.1% 的 SNP 位点的 p 值得观测分布与 p 值的期望分布一致，当 SNP 位点的 p 值小于 10^{-2}时，QQ 图尾巴出现上翘且上翘结果理想。除此之外，该模型还在 0.2 Mbp 和 5.3 Mbp 附近检测出了 2 个新的峰区，说明 $\boldsymbol{Q}+\boldsymbol{K}$ 模型对 SNP 位点的筛查能力更强。

对功能作图模型、一般随机效应模型、$\boldsymbol{Q}+\boldsymbol{K}$ 混合效应功能作图模型检测出的 SNP 进行基因功能注释（表 2），以基因 *ruvB*、*mdtM*、*htpX* 和 *gldA* 为例。通过注释信息，发现基因 *ruvB* 参与大肠杆菌的基因修复过程，有助于大肠杆菌的繁殖；基因 *mdtM* 可参与大肠杆菌的排出毒性化合物过程；基因 *htpX* 有助于大肠杆菌蛋白质活性的提高；基因 *gldA* 通过

表 2 显著位点基因注释信息

Tab. 2 Gene annotation information with significant SNPs

模型 Model	SNP编号 SNP No.	基因 Gene	注释 Annotation
功能作图 Functional mapping	19513	*ydjL*	假定的锌结合蛋白 Putative, zinc-binding
功能作图 Functional mapping	20057	*htpX*	锌依赖性内切蛋白酶 Zinc, dependent, endoprotease
一般随机效应 General random effect	14587	*trpB*	色氨酸合酶β亚基 Tryptophan, synthase, subunit, beta
一般随机效应 General random effect	19513	*ydjL*	假定的锌结合蛋白 Putative, zinc-binding
一般随机效应 General random effect	20057	*htpX*	锌依赖性内切蛋白酶 Zinc, dependent, endoprotease
一般随机效应 General random effect	20249	*yobB*	假定的碳氮水解酶家族蛋白 YobB Putative, carbon-nitrogen, hydrolase family protein YobB
一般随机效应 General random effect	44357	*nikB*	Ni_2^+ABC细胞膜转运体亚基 Putative, carbon-nitrogen, hydrolase family protein YobB
一般随机效应 General random effect	50949	*gldA*	L-1, 2-丙二醇脱氢酶/甘油脱氢酶 L-1, 2-propanediol, dehydrogenase/glycerol dehydrogenase
一般随机效应 General random effect	51244	*ppc*	磷酸烯醇式丙酮酸碳羧化酶 Phosphoenolpyruvate carboxylase
混合效应功能作图 Mixed effect functional mapping	14587	*trpB*	色氨酸合酶β亚基 Tryptophan, synthase, subunit, beta
混合效应功能作图 Mixed effect functional mapping	19513	*ydjL*	假定的锌结合蛋白 Putative, zinc-binding
混合效应功能作图 Mixed effect functional mapping	20057	*htpX*	锌依赖性内切蛋白酶 Zinc, dependent, endoprotease
混合效应功能作图 Mixed effect functional mapping	20249	*yobB*	假定的碳氮水解酶家族蛋白 YobB Putative, carbon-nitrogen, hydrolase family protein YobB
混合效应功能作图 Mixed effect functional mapping	20285	*ptrB*	寡肽酶 B Oligopeptidase B
混合效应功能作图 Mixed effect functional mapping	20455	*ruvB*	RuvAB-霍利迪连接体分支迁移复合物亚基 Holliday, junction branch migration complex subunit RuvB
混合效应功能作图 Mixed effect functional mapping	20794	*argS*	精氨酸-tRNA 连接酶 Arginine--tRNA ligase
混合效应功能作图 Mixed effect functional mapping	44357	*nikB*	Ni_2^+ABC细胞膜转运体亚基 Ni_2^+ABC, transporter membrane subunit NikB
混合效应功能作图 Mixed effect functional mapping	49473	*metE*	不依赖钴胺素的同型半胱氨酸转甲基化酶 Cobalamin-independent, homocysteine transmethylase
混合效应功能作图 Mixed effect functional mapping	49950	*glnL*	组氨酸激酶 NtrB Sensory histidine kinase NtrB
混合效应功能作图 Mixed effect functional mapping	50949	*gldA*	L-1, 2-丙二醇脱氢酶/甘油脱氢酶 L-1, 2-propanediol, dehydrogenase/glycerol dehydrogenase
混合效应功能作图 Mixed effect functional mapping	51244	*ppc*	磷酸烯醇式丙酮酸碳羧化酶 Phosphoenolpyruvate carboxylase
混合效应功能作图 Mixed effect functional mapping	57014	*mdtM*	多药外排泵/胆酸盐 H^+拮抗剂/Na^+:H^+拮抗剂/K^+: H^+拮抗剂 Multidrug, efflux, pump/bile, salt: H^+ antiporter/Na^+:H(+)antiporter/K^+: H^+antiporter
混合效应功能作图 Mixed effect functional mapping	57227	*opgB*	磷酸甘油基 Phosphoglycerol
混合效应功能作图 Mixed effect functional mapping	57749	*deoC*	脱氧核糖磷酸 Deoxyribose-phosphate

催化反向反应来调节大肠杆菌二羟基丙酮的水平，对大肠杆菌的生长发挥着不可或缺的作用。功能注释结果表明，本文采用 $\boldsymbol{Q}+\boldsymbol{K}$ 模型对使用以逻辑斯谛曲线为例的功能作图具有较好的定位效果，进一步证明了本文在功能作图领域所拓展的混合效应模型的合理性。

3.5 计算机模拟

表 3 展示了不考虑混合效应的功能作图模型和混合效应功能作图模型在不同遗传力和不同样本量组合情况下的模型功效。表 3 显示：功能作图模型在样本量较小、遗传力较低时，模型假阳性率是 35.71%；而混合效应功能作图模型在样本量较小、遗传力较低时，模型假阳性率约为 4.77%，模型假阳性率显著下降，不同样本量的假阳性率都在 5% 以下。另外，在遗传力为 0.25% 和 0.50% 时，功能作图模型的模型功效高于混合效应功能作图模型，考虑前者在遗传力较低时出现了 35.71% 的假阳性率，这有可能是功能作图模型出现高假阳性率的结果。随着样本量和遗传力的不断提升，当样本量为 800，遗传力为 1% 时，或当样本量为 400，遗传力为 1.5% 时，功能作图模型和混合效应功能作图模型具有相近的模型功效。当样本量为 800，遗传力为 1.5% 时，或当样本量为 400，遗传力为 2% 时，混合效应功能作图模型的模型功效更高，QTL 发现率超过 70%。从表

表 3　2 种模型在不同遗传力和不同样本量组合下的模型功效

Tab. 3　Model power for different combinations of heritability and sample size

样本量 Sample number	模型 Model	遗传力 Heritability/%					
		0.0	0.25	0.50	1.0	1.5	2.0
100	$\boldsymbol{Q}$	0.357 1	0.374 6	0.438 6	0.512 3	0.596 3	0.675 1
	$\boldsymbol{Q}+\boldsymbol{K}$	0.047 7	0.072 2	0.107 5	0.186 9	0.287 6	0.397 1
200	$\boldsymbol{Q}$	0.348 6	0.396 8	0.485 6	0.554 2	0.651 5	0.701 5
	$\boldsymbol{Q}+\boldsymbol{K}$	0.047 0	0.097 6	0.165 9	0.334 4	0.507 9	0.667 8
400	$\boldsymbol{Q}$	0.308 2	0.434 6	0.538 7	0.611 4	0.728 5	0.785 6
	$\boldsymbol{Q}+\boldsymbol{K}$	0.038 5	0.119 7	0.232 6	0.495 0	0.733 9	0.882 3
600	$\boldsymbol{Q}$	0.296 3	0.448 9	0.587 4	0.668 1	0.804 2	0.869 6
	$\boldsymbol{Q}+\boldsymbol{K}$	0.030 0	0.131 9	0.278 6	0.627 2	0.847 5	0.945 4
800	$\boldsymbol{Q}$	0.232 8	0.466 2	0.627 5	0.695 1	0.886 2	0.906 1
	$\boldsymbol{Q}+\boldsymbol{K}$	0.019 6	0.136 1	0.321 9	0.705 3	0.902 3	0.964 8

注: $\boldsymbol{Q}$. 功能作图模型; $\boldsymbol{Q}+\boldsymbol{K}$. 混合效应功能作图模型。Notes: $\boldsymbol{Q}$, functional mapping; $\boldsymbol{Q}+\boldsymbol{K}$, mixed effect functional mapping.

中结果得知: 与功能作图模型相比, 混合效应功能作图模型具有更低的假阳性率和更高的 QTL 发现率。

4　结论与讨论

功能作图是一项针对动态性状的 QTL 定位技术, 其核心是使用生长曲线与协方差模型的方式对动态表型进行合理利用与分析, 使 QTL 定位在遗传与统计相结合的基础上更具生物学内涵。本研究从最经典的功能作图技术入手, 从固定效应和随机效应的准确剖分角度出发, 提出了功能作图的 $\boldsymbol{Q}$ 模型扩展; 基于限制性最大似然估计的统计框架, 构建了功能作图的非线性混合效应模型; 提出 Zwald 检验方法; 并以大肠杆菌的动态生长速率为实际数据, 应用逻辑斯谛生长曲线与 LOP 的非参数曲线对上述模型开展实例分析, 验证了模型的可行性。

首先, 基于功能作图所采用的生长曲线, 本研究对该模型的生长曲线参数作了固定效应的 $\boldsymbol{Q}$ 模型扩展, 认为不同因子对动态表型的作用方式体现于生长曲线参数的分解, 为 QTL 效应的准确剖分奠定了基础, 这在大肠杆菌例子中的 11 个亚群效应上得到了较好体现。对于本文利用的 11 个亚群, 其 $\boldsymbol{Q}$ 矩阵表现为各菌株归属于 11 个亚群的概率, 累加和为 1, 对 SNP 的列信息则由 0 或 1 填充, 因此合并所得的 $\boldsymbol{Q}$ 矩阵既包含了连续变量又包含了分类变量。在定位群体的背景较为复杂, 或群体需考虑的协变量因素较多的情况下, 该扩展可有效避免功能作图在各时间点上依次对协变量开展回归和校正的步骤, 并且可以使各 SNP 位点上的分析体现出不同的协变量效应, 更有利于 SNP 效应的准确估算。

其次, 早期经典的功能作图技术考虑了重复测量表型之间的自相关关系, 对此采用协方差结构建模。Hou 等参考文献附上在此基础上新增了个体之间因异质性而产生的随机效应, 并将随机效应写成了生长曲线参数的形式, 认为随机效应也是通过生长曲线参数的方式影响动态生长过程。经本文推导, 该处理方式对模型 1, 即一般化随机效应的混合模型具有计算可行的结果, 但对 $\boldsymbol{Q}+\boldsymbol{K}$ 混合模型不具备求解结果。对此, 本研究采用 LOP 的非参曲线单独对随机效应建模, 该剖分方式可使随机效应的形式更加固定化, 更利于功能作图在不同数据场合下的应用。

第三, 本研究延用 Hou 等用一阶泰勒展开的方式将 NLME 转化为 LMM, 该方法为混合模型的求解提供了便利。传统功能作图使用数值解的方法需要对方差–协方差参数进行求解, 同时仍要对曲线参数进行数值解的最优搜索。本研究通过提出 REML 算法, 与最大似然估计相比, 可求出方差–协方差参数的无偏估计值; 并且在一节泰勒展开的线性化 LMM 框架下, 借助方差–协方差参数的无偏估计结果, 关于 $\boldsymbol{Q}$ 模型中的生长曲线参数可通过(18)式直接计算得到, 较传统方法减少了参数求解个数, 使功能作图在统计计算上具备了简约可行的特点。基于 REML 算法, 进一步构建的 Zwald 统计量可作为似然比检验的另一个补充, 方便用户选用合适的统计量检测筛选出合适的 QTL。

QTL 定位是开展复杂性状遗传解析的关键技术之一, 自 1989 年 Lander 与 Bostein[29]开启了数量性状的 QTL 定位技术之旅以来, 人们就开始发展不同的统计模型来研究数量性状。直到 Yu 等[16]基于常规静态性状, 开发了复杂的 $\boldsymbol{Q}+\boldsymbol{K}$ 模型。自此之后, 有不同的 $\boldsymbol{Q}+\boldsymbol{K}$ 算法被提出[30−32], 用以解决群体的混杂协变量和由亲缘关系所决定的个体间相关性造成的随机效应。本研究瞄准动态性状的 $\boldsymbol{Q}+\boldsymbol{K}$ 模型, 以逻辑斯谛曲线的生长规律为例, 发展了 2 类混合效应模型。联合 Zwald 检验的统计方法表明, $\boldsymbol{Q}+\boldsymbol{K}$ 模型对动态性状的 QTL 定位可起到较好的校正作用。在动物育种领域, Ning 等[33]以 LOP 作为通用的生长曲线, 开发了 $\boldsymbol{Q}+\boldsymbol{K}$ 的测试日模型。LOP 曲线

是一类非参数估计方法，与此相比，本文所采用的逻辑斯谛生长曲线更具生物学内涵。比如，基于曲线求导，生长关键时间点、快速生长阶段、渐近生长时间点都可被有效计算。因而，扩充更多的生长曲线来进一步发展功能作图的 $\boldsymbol{Q}+\boldsymbol{K}$ 模型，将成为功能作图的下一步研究方向。与静态性状相比，动态性状的 QTL 定位技术虽然起步较晚，但可以肯定的是，联合高通量标记和高维表型数据的功能作图 $\boldsymbol{Q}+\boldsymbol{K}$ 模型将对未来复杂性状的解析起到变革性的作用。

参 考 文 献

[1] Das K, Li J, Fu G, et al. Dynamic semiparametric Bayesian models for genetic mapping of complex trait with irregular longitudinal data[J]. Statistics in Medicine, 2013, 32(3): 509−523.

[2] Sun L D, Wu R L. Mapping complex traits as a dynamic system[J]. Physics of Life Reviews, 2015, 13: 155−185.

[3] Ma C X, Casella G, Wu R. Functional mapping of quantitative trait loci underlying the character process: a theoretical framework[J]. Genetics, 2002, 161(4): 1751−1762.

[4] 甘静雯. 基于随机微分方程的复杂性状功能作图模型构建 [D]. 北京: 北京林业大学, 2021.
Gan J W. Construction of functional mapping model for complex traits based on stochastic differential equations[D]. Beijing: Beijing Forestry University, 2021.

[5] Lou X, Yang M, Wu R L, et al. A general statistical framework for unifying interval and linkage disequilibrium mapping: towards highresolution mapping of quantitative traits[J]. Journal of the American Statistical Association, 2005, 100: 158−171.

[6] Wu R L, Ma C X, Lin M, et al. A general framework for analyzing the genetic architecture of developmental characteristics[J]. Genetics, 2004, 166(3): 1541−1551.

[7] Wu R L, Lin M. Functional mapping: how to map and study the genetic architecture of dynamic complex traits[J]. Nature Reviews Genetics, 2006, 7(3): 229−237.

[8] Chitwood D H, Topp C N. Revealing plant cryptotypes: defining meaningful phenotypes among infinite traits[J]. Current Opinion in Plant Biology, 2015, 24: 54−60.

[9] Li Z, Sillanpaa M J. Dynamic quantitative trait locus analysis of plant phenomic data[J]. Trends in Plant Science, 2015, 20(12): 822−833.

[10] Ye M X, Jiang L B, Mao K, et al. Functional mapping of seasonal transition in perennial plants[J]. Briefings in Bioinformatics, 2015, 16(3): 526−535.

[11] Zhao W, Hou W, Littell R C, et al. Structured antedependence models for functional mapping of multiple longitudinal traits[J]. Statistical Applications in Genetics and Molecular Biology, 2005, 4: e33.

[12] Wu R L, Ma C X, Littell R C, et al. A logistic mixture model for characterizing genetic determinants causing differentiation in growth trajectories[J]. Genetics Resesearch, 2002, 79(3): 235−245.

[13] Hou W, Li H, Zhang B, et al. A nonlinear mixed-effect mixture model for functional mapping of dynamic traits[J]. Heredity (Edinb), 2008, 101(4): 321−328.

[14] Zhu H, Lee S. Analysis of generalized linear mixed models via a stochastic approximation algorithm with Markov chain Monte-Carlo method[J]. Statistics and Computing, 2002, 12(2): 175−183.

[15] Stoica P, Moses R, Friedlander B, et al. Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989, 37(3): 379−392.

[16] Yu J, Pressoir G, Briggs W H, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness[J]. Nature Genetics, 2006, 38(2): 203−208.

[17] Meyer K. WOMBAT: a tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML)[J]. Journal of Zhejiang University-Science B, 2007, 8(11): 815−821.

[18] Hauck W W, Donner A. Corrigenda: “Wald’s test as applied to hypotheses in logit analysis”[J]. Journal of the American Statistical Association, 1980(370): 482.

[19] Marchelan F, Osilenker B P. Estimates for Legendre-Sobolev polynomials that are orthogonal with respect to the scalar product[J]. Journal of Inequalities and Applications, 1997(6): 871−880.

[20] Zhao W, Chen Y Q, Casella G, et al. A non-stationary model for functional mapping of complex traits[J]. Bioinformatics, 2005, 21(10): 2469−2477.

[21] Nelder J A, Mead R. A simplex method for function minimization[J]. Computer Science, 1965, 7: 308−313.

[22] Lindstrom M J, Bates D M. Correction to: “Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures-data”[J]. Journal of the American Statistical Association, 1994(428): 1572.

[23] Laird N M, Ware J H. Random-effects models for longitudinal data[J]. Biometrics, 1982, 38(4): 963−974.

[24] Vanraden P M. Efficient methods to compute genomic predictions[J]. Journal of Dairy Science, 2008, 91(11): 4414−4423.

[25] Yang D, Jin Y, He X, et al. Inferring multilayer interactome networks shaping phenotypic plasticity and evolution[J]. Nature Communications, 2021, 12(1): 5304.

[26] Jiang L, He X, Jin Y, et al. A mapping framework of competition-cooperation QTLs that drive community dynamics[J]. Nature Communications, 2018, 9(1): 3010.

[27] Raj A, Stephens M, Pritchard J K. fastSTRUCTURE: variational inference of population structure in large SNP data sets[J]. Genetics, 2014, 197(2): 573−589.

[28] Jiang L, Clavijo J A, Sun L, et al. Plastic expression of heterochrony quantitative trait loci (hQTLs) for leaf growth in the common bean (*Phaseolus vulgaris*)[J]. New Phytologist, 2015, 207(3): 872−882.

[29] Lander E S, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps[J]. Genetics, 1989, 121(1): 185−199.

[30] Kang H M, Zaitlen N A, Wade C M, et al. Efficient control of population structure in model organism association mapping[J]. Genetics, 2008, 178(3): 1709−1723.

[31] Zhang Z, Ersoz E, Lai C Q, et al. Mixed linear model approach adapted for genome-wide association studies[J]. Nature Genetics, 2010, 42(4): 355−360.

[32] Lippert C, Listgarten J, Liu Y, et al. Fast linear mixed models for genome-wide association studies[J]. Nature Methods, 2011, 8(10): 833−835.

[33] Ning C, Wang D, Zhou L, et al. Efficient multivariate analysis algorithms for longitudinal genome-wide association studies[J]. Bioinformatics, 2019, 35(23): 4879−4885.

（责任编辑 范 娟 赵田芸
责任编委 康向阳）