遗传 Hereditas (Beijing) 2021 年 4 月, 43(4): 340－349
www.chinagene.cn

研究报告

基于高密度 SNP 标记估计群体间遗传关联

周子文，王雪，丁向东

中国农业大学动物科技学院，畜禽育种国家工程实验室，农业农村部动物遗传育种与繁殖重点实验室，北京 100193

摘要: 联合育种的准确性受到群体间遗传关联程度的影响。本研究通过比较基于系谱数据和基因组数据计算的群体遗传关联，探究高密度 SNP 标记在遗传关联估计中的应用前景。本研究同时使用了模拟数据和真实数据，采用 6 种不同的遗传关联计算方法，包括 PEVD (prediction error variance of differences)、PEVD(x)、VED (variance of estimated difference)、CD (generalized coefficient of determination)、r (prediction error correlation)和 CR (connectedness rating)，比较基于构建不同的关系矩阵(A、G、G_s、$G_{0.5}$和 H 矩阵)的群体间遗传关联。模拟数据和实际数据结果表明，除 PEVD(x)和 VED 方法外，PEVD、CD、r 和 CR 基于基因组信息的 G、G_s和 $G_{0.5}$阵计算的遗传关联程度均高于基于系谱信息的 A 阵，基于同时利用系谱和基因组信息的 H 阵遗传关联结果一般介于 A 阵与 G 阵之间。当 CR 和 r 为 0 时，CD 都较高，高估了群体遗传关联。用 r 度量 3 个遗传分化程度不同的猪场间遗传关联时，基于 G 阵的 r 值均为 0.01，不能准确反映群体真实遗传关联。随着遗传力的提高，所有群体遗传关联评估方法都有所改善，但遗传力为 0.1 时，PEVD 基于 A 阵结果优于 G 阵，中高遗传力性状用于估计遗传关联优于低遗传力性状。本研究证明高密度 SNP 标记比系谱信息估计群体间遗传关联更有优势，CR 是衡量遗传关联稳健而可靠的评价指标，计算简单，受性状遗传力影响较小。PEVD 可以作为补充，量化具体群体遗传关联下的育种值预测误差情况。G 矩阵比 G_s、$G_{0.5}$阵能更好反映群体遗传关联。

关键词: 猪；遗传关联；系谱；基因组；关系矩阵

Measuring genetic connectedness between herds based on high density SNP markers

Ziwen Zhou, Xue Wang, Xiangdong Ding

National Engineering Laboratory for Animal Breeding, Key Laboratory of Animal Genetics, Breeding and Reproduction of Ministry of Agriculture and Rural Affairs, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China

Abstract: The accuracy of genetic evaluations in different herds is affected by the degree of genetic connectedness among herds. In this study, we explored the application of high density SNP markers in the assessment of genetic

收稿日期: 2020−10−19; **修回日期:** 2021−02−17

基金项目: 国家现代农业产业技术体系项目(编号：CARS-35)，国家重点研发计划项目(编号：2019YFE0106800)和河北省重点研发计划项目(编号：19226376D)资助[Supported by China Agriculture Research System (No. CARS-35), the National Key Research and Development Project (No. 2019YFE0106800) and Modern Agriculture Science and Technology Key Project of Hebei Province (No. 19226376D)]

作者简介: 周子文，在读硕士研究生，专业方向：动物遗传育种。E-mail: zhouzw834@163.com

通讯作者: 丁向东，博士，副教授，研究方向：猪遗传育种和统计遗传学。E-mail: xding@cau.edu.cn

DOI: 10.16288/j.yczz.20-351

网络出版时间: 2021/3/29 16:45:30

URI: https://kns.cnki.net/kcms/detail/11.1913.r.20210326.1346.004.html

connectedness by comparing the genetic connectedness based on pedigree data and genomic data. Six methods, including PEVD (prediction error variance of differences between estimated breeding values), PEVD (x), VED (variance of estimated difference between the herd effects), CD (generalized coefficient of determination), r (prediction error correlation) and CR (connectedness rating), were implemented to measure the genetic connectedness based on different relationship matrices (A, G, G_s, $G_{0.5}$ and H). Our results from both simulated data and SNP chip data indicated that, except for the PEVD (x) and VED methods, the genetic connectedness obtained by PEVD, CD, r and CR based on G. G_s and $G_{0.5}$ matrices (using genome information only) were superior to those based on A matrix (using pedigree information only). Generally, for most approaches, the genetic connectedness based on H matrix (using both pedigree and genome information) was somewhere between A matrix and G matrices. CD could overestimate the degree of genetic connectedness as it was still very high when CR and r were close to 0. The method r could not accurately reflect the true genetic connectedness of the populations. It generated 0.01 of genetic connectedness for all three pig breeding farms, which were actually genetically different with each other. With increasing of heritability, the degree of genetic connectedness obtained by all methods were increased as well. However, in the case of heritability 0.1, PEVD based on A matrix performed better than based on G matrix, suggesting that traits with medium and high heritability are more suitable for the assessment of genetic connectedness compared to traits with low heritability. Our findings indicated that high-density SNP markers have advantages over pedigree analysis for the measurement of genetic connectedness, and CR is a robust and reliable method to assess genetic connectedness. Further, CR is easily calculated and less affected by heritability of trait. PEVD is good supplement to quantify the prediction errors of estimated breeding values under the specific genetic connectedness. In comparison, G matrix can reflect genetic connectedness better than its extensions G_s and $G_{0.5}$ matrix.

Keywords: swine; genetic connectedness; pedigree; genome; relationship matrix

联合评估是家畜育种中一种有效方式，可将不同国家、地区、育种组织的育种数据联合起来对个体进行遗传评估，由于扩大了群体规模，从而提高了育种值估计的准确性，而且能够进行大范围内种畜的比较和选种，实现联合育种。奶牛、猪育种中这一做法十分通行，但多个群体的联合遗传评估，群体间存在关联是前提，表现为群体具有遗传上的关联或者由于相同环境造成的关联[1]，从而使多个群体的联合评估可以在同一尺度上进行比较。相同环境所造成的群体关联主要通过不同群体在中心测定站统一进行性能测定实现，但由于中心测定站测定规模限制，此种群体关联影响有限。群体关联更多是由于场间遗传交流产生的遗传联系，如我国生猪遗传改良计划开展的联合遗传评估，通过场间遗传交流建立不同生猪核心育种场之间的遗传联系，形成了杜洛克、长白和大白 3 个品种多个遗传关联组[2]。每个关联组联合评估与单场遗传评估相比，遗传参数估计和个体育种值估计准确性更高[3]，并且可以进行个体跨场比较，挑选优秀种猪。

群体遗传关联有多种估计方法，可以分为两大类：育种值估计预测误差方差和育种值比较的可靠性或相关系数。第一种主要有预测误差方差方法(prediction error variance of differences, PEVD)[4]、PEVD(x)[5]和场效应差异的估计方差(variance of estimated difference, VED)[4]。从理论上说，PEVD 是一种较为理想的度量遗传关联的方法，该方法通过计算不同个体之间育种值(estimated breeding value, EBV)差异的预测误差方差，评价两个个体育种值比较的准确性，但该方法计算复杂，难以用于育种实践[4]。PEVD(x)和 VED 是 PEVD 的近似估计方法，PEVD(x)通过构建一个差异向量 x 近似估计 PEVD，进行简化计算[5]，VED 主要计算场效应之间的预测误差方差[4]。第二类群体遗传关联估计方法主要有广义决定系数方法(generalized coefficient of determination, CD)[5]、预测误差相关系数(prediction error correlation, r)[6]和场间关联率(connectedness

rating, CR)[7]。CD 定义为估计育种值比较的可靠性，即预测值差异与真实值差异间相关系数的平方[5]，r 通过计算两个群体之间两两配对的预测误差相关系数均值来评价遗传关联程度[6]。CR 主要计算场效应之间的相关，或者群体均值估计误差之间的相关[7]。

遗传关联计算通常基于系谱数据[8]，但是系谱数据难以保证其完整性和准确性，会导致部分场间遗传关联低于真实情况，或者产生错误的场间遗传关联。如果两个群体均有基因组数据，则即便缺乏完整准确的系谱记录，也可以估计遗传关联，从而拓展了遗传关联方法的使用范围。使用基因组数据估计遗传关联的另一个主要优势为基因组数据能够真实反映个体间亲缘关系，通过基因组数据构建的个体间关系矩阵比基于系谱信息的更加准确[9~11]，可以捕捉到系谱数据中不存在的遗传关联。

本研究旨在通过比较不同群体关联估计方法基于系谱和 SNP 芯片数据计算的遗传关联，探究基因组数据在遗传关联估计中的应用效果及各种群体关联估计方法的优劣。

1 数据与方法

1.1 模拟数据

本研究采用 GPOPSIM[12]软件模拟基因组数据。模拟了 18 条染色体，每条染色体长度为 100 cM，染色体总长度为 18 M，总共模拟了 306 个 QTL，随机分布在染色体组上。SNP 标记和 QTL 的突变率分别为 1.25×10^{-6} 和 2.5×10^{-3}。从每条染色体上均匀抽取 2834 个 SNP，共 51,012 个 SNP，生成基因型数据。表型数据由软件模拟生成，遗传力设定为 0.3，遗传方差为 2。

群体模拟首先生成一个 1000 世代的历史群体，每个世代群体规模保持不变，均由 300 头公畜和 300 头母畜组成，公母随机交配，每头母畜产生 10 个后代，公母各半。从第 1000 个世代群体后代中随机抽取，生成两个亚群，每个亚群均由 20 头公畜和 600 头母畜构成。每个亚群内，每头公畜与 30 头母畜随机交配，每头母畜产生 10 个后代，公母比例 1∶1，记为世代 1。从世代 2 开始，每个亚群内均从上一世代随机选择 20 头公畜与 1500 头母畜交配，母畜产生后代数与性别比例同世代 1，不同世代群体大小保持不变。重复上述过程，直至世代 7，两个亚群间不发生遗传交流。两个亚群世代 1 至世代 7 所有个体均有表型，仅第 5 世代至第 7 世代每个亚群各有 3000 个体(每个公畜家系中一半个体)具有基因型数据。

1.2 真实数据

本研究同时利用 3 家国家生猪核心育种场(以下简称“核心场”)北京六马养猪科技股份有限公司(场代码 BJLM，简称“北京六马”)、北京养猪育种中心(场代码 BBSC，简称“养猪中心”)及新疆天康畜牧科技有限公司(场代码 XJTC，简称“新疆天康”) 2012~2019 年大白猪数据。北京六马和养猪中心种猪来源于美国，新疆天康来源于加拿大。3 家核心场生长性状达 100 kg 体重日龄和 100 kg 活体背膘厚表型数据分别为 33,883、13,259 和 13,763 条，系谱数据各有 36,577、75,255 和 14,409 条，具有 SNP 芯片基因型个体数为 2382、1712 和 1239 头。

北京六马和养猪中心的基因型数据均采用 PorcineSNP80K Beadchip 芯片(简称 80K)测定，共包含 68,528 个 SNP 位点；新疆天康的基因型数据则由 PorcineSNP50K Beadchip 芯片(简称 50K)测定，共包含 50,697 个 SNP 位点。两种芯片均参照猪参考基因组 11.1 版本，除去未知染色体上的位点后，两款芯片共同位点数为 48,675。芯片基因型填充步骤分两步进行，首先对 80K 芯片个体进行填充，剔除未知染色体和常染色体上的位点，之后将其作为参考群对所有 50K 芯片个体进行填充，芯片数据填充处理使用 beagle[13]软件完成。填充后对芯片数据进行如下质控处理：(1)个体检出率(call rate)达到 90%以上；(2)单个 SNP 检出率达到 90%以上；(3) SNP 位点的最小等位基因频率不低于 0.05；(4)每个 SNP 位点哈代–温伯格平衡检验 P 值大于 10^{-6}。质控筛选后，所有基因型个体、45569 个 SNP 位点满足要求。

1.3 育种值估计模型

群体关联估计主要基于育种值估计，本研究所有群体遗传关联方法计算均基于以下育种值估计模型：

$$y = \mathrm{X}b + \mathrm{Z}a + e$$

其中，$\boldsymbol{y}$ 为单一性状表型值向量(实际数据用达100kg 体重日龄)；$\boldsymbol{b}$ 为场效应向量；$\boldsymbol{a}$ 为加性遗传随机效应或育种值向量，服从正态分布 $N(0, \mathbf{K}\sigma_a^2)$，$\sigma_a^2$ 为加性遗传方差，**K** 阵为亲缘关系矩阵，可使用 **A** 阵、**G** 阵、**H** 阵等，主要基于系谱信息、基因组信息或同时利用系谱–基因组信息建立；$\boldsymbol{e}$ 为随机残差，服从正态分布 $N(0, \mathbf{I}\boldsymbol{\sigma_e^2})$，$\boldsymbol{\sigma_e^2}$ 为残差方差；**X** 和 **Z** 为相应的结构矩阵。

1.4　遗传关联计算方法

本研究使用 PEVD、PEVD(x)、VED、CD、r 和 CR 等 6 种方法估计群体遗传关联。PEVD 计算公式如下：

$$\mathrm{PEVD}(\hat{\mu}_i - \hat{\mu}_j) = [\mathrm{PEV}(\hat{\mu}_i) + \mathrm{PEV}(\hat{\mu}_j) - 2\mathrm{PEC}(\hat{\mu}_i, \hat{\mu}_j)] = (\mathrm{C}_{ii}^{22} + \mathrm{C}_{jj}^{22} - 2\mathrm{C}_{ij}^{22})\sigma_e^2$$

其中 $\mathrm{PEV}(\hat{u}_i)$、$\mathrm{PEV}(\hat{u}_j)$ 和 $\mathrm{PEC}(\hat{u}_i, \hat{u}_j)$ 分别为个体加性效应的预测误差方差和协方差，$\mathbf{C}^{22}$ 矩阵为系数矩阵逆矩阵的子矩阵，$\boldsymbol{\sigma_e^2}$ 为残差方差。以上参数均可通过育种值估计模型求解获得。群体间 PEVD 通过计算属于不同群体的所有个体两两配对的 PEVD 均值得到；群体内 PEVD 通过计算单个群体内的所有个体两两配对的 PEVD 均值得到。

PEVD(x)方法参照 Laloë 等[5]。VED、CD、r 和 CR 方法计算公式如下：

$$\mathrm{VED} = \mathrm{var}(\hat{\mathrm{h}}_1) + \mathrm{var}(\hat{\mathrm{h}}_2) - 2\,\mathrm{cov}(\hat{\mathrm{h}}_1, \hat{\mathrm{h}}_2)$$

$$\mathrm{CD}_{ij} = 1 - \lambda\frac{\mathbf{C}_{ii}^{22} + \mathbf{C}_{jj}^{22} - 2\mathbf{C}_{ij}^{22}}{\mathbf{K}_{ii} + \mathbf{K}_{jj} - 2\mathbf{K}_{ij}}$$

$$\mathrm{r}_{ij} = \frac{\mathrm{PEC}(\hat{u}_i, \hat{u}_j)}{\sqrt{\mathrm{PEV}(\hat{u}_i)\,\mathrm{PEV}(\hat{u}_j)}}$$

$$\mathrm{CR} = \frac{\mathrm{cov}(\hat{\mathrm{h}}_1, \hat{\mathrm{h}}_2)}{\sqrt{\mathrm{var}(\hat{\mathrm{h}}_1)\,\mathrm{var}(\hat{\mathrm{h}}_2)}}$$

其中 $\mathbf{K}_{ii}$、$\mathbf{K}_{jj}$ 和 $\mathbf{K}_{ij}$ 为关系矩阵中的相应元素，λ 为随机残差与加性方差的比值；$\mathrm{var}(\hat{\mathrm{h}}_1)$、$\mathrm{var}(\hat{\mathrm{h}}_2)$ 和 $\mathrm{cov}(\hat{\mathrm{h}}_1, \hat{\mathrm{h}}_2)$ 分别为场效应的方差和协方差。与 PEVD 类似，群体间的 CD 和 r 计算方法为群体间所有个体两两匹配求均值。本研究中所有遗传关联方法计算均通过自编 R 程序实现。6 种计算方法中，CD、r 和 CR 范围在 0~1 之间，当 r 或 CR 为 0 时，说明两个群体之间缺乏遗传关联，CD、r 和 CR 值越大说明遗传关联程度越高。PEVD、PEVD(x)和 VED 值越小，说明遗传关联程度越高。

1.5　亲缘关系矩阵构建

本研究主要利用系谱数据和 SNP 芯片数据，构建 1.3 中的 K 阵，包括 **A** 阵、**G** 阵、$\mathbf{G_s}$阵、$\mathbf{G_{0.5}}$阵和 **H** 阵。**G** 阵构建参照 VanRaden 等[14]，为防止 **G** 阵为奇异阵导致无法求逆，本研究中将 **G** 阵对角线元素均增加了 0.01[15]。研究表明，使用 **G** 阵计算预测误差相关系数 r 时，r 容易出现负值[16,17]。本研究通过将 **G** 阵中的负值替换为 0，避免了 r 和 CR 方法结果出现负值。**G** 阵构建中需要每个标记在基础群体时的最小等位基因频率，通常用当前群体的最小等位基因频率代替，$\mathbf{G_{0.5}}$ 矩阵将最小等位基因频率均假设为 0.5[18,19]。

为校正 **G** 阵中元素大小，使 **G** 阵与 **A** 阵尺度保持一致。本研究将 **G** 阵中的所有元素校正到给定的最小值和最大值的范围之内，命名为 $\mathbf{G_s}$ 阵。$\mathbf{G_s}$ 矩阵构建方法如下：

$$\mathbf{G}_{\mathbf{s}_{ij}} = \frac{(\mathbf{G}_{\mathrm{smax}} - \mathbf{G}_{\mathrm{smin}})(\mathbf{G}_{\mathrm{ij}} - \mathbf{G}_{\mathrm{min}})}{\mathbf{G}_{\mathrm{max}} - \mathbf{G}_{\mathrm{min}}}$$

其中 $\mathrm{G_{smax}}$、$\mathrm{G_{smin}}$ 为给定的 $\mathbf{G_s}$ 矩阵的最大值和最小值，本研究分别设定为 2 和 0，以模拟 **A** 阵中的最大值和最小值；$\mathrm{G_{max}}$、$\mathrm{G_{min}}$ 为 G 阵中的最大值和最小值；$\mathrm{G_{ij}}$ 为相应的 **G** 阵元素。

本研究 H 阵构建参照 Legarra 等[20]，其中 **H** 阵中使用的 **G** 阵经过了两步校正，第一步校正参照 Legarra 等[20]，生成一个新的矩阵 $\mathbf{G_a}$，保证 **G** 阵和 **A** 阵具有相同的尺度。由于基因型数据不能完全解释基因组信息，赋予 $\mathbf{G_a}$ 阵和 **A** 阵不同的权重，生成新的 $\mathbf{G_w}$ 阵，最终用于 **H** 阵构建。本研究 $\mathbf{G_a}$ 阵和 **A** 阵权重分别为 0.95 和 0.05。

2　结果与分析

2.1　群体关联估计方法和关系矩阵影响

表 1 反映了基于模拟数据，6 种群体关联估计方法和 5 种关系矩阵对群体关联估计的影响。以模拟数据第 5 世代两个亚群群体关联结果为例，使用

G 阵相较于 **A** 阵能够提高群体遗传关联。PEVD 从 1.65 降至 1.32，$\mathbf{G_{0.5}}$ 阵则进一步使 PEVD 降低至 0.9285。基于 $\mathbf{G_s}$ 阵估计的 PEVD 高于 **G** 阵，但仍低于 A 阵，同时利用系谱和基因组信息的 **H** 阵 PEVD 与 **G** 阵接近。作为 PEVD 的扩展，PEVD(x)和 VED 方法却呈现了相反趋势，**G**、$\mathbf{G_s}$、$\mathbf{G_{0.5}}$ 阵结果劣于 **A** 阵，基于 **A** 阵的 PEVD(x)和 VED 过低，接近于 0。由于受 A 阵影响，基于 **H** 阵的 PEVD(x)和 VED 也很小，分别为 0.002 和 0.004。**G**、$\mathbf{G_s}$、$\mathbf{G_{0.5}}$ 矩阵 PEVD(x)和 VED 在 0.27~0.42 间变化，$\mathbf{G_{0.5}}$ 最小，$\mathbf{G_s}$ 最大。

对于 r 和 CR，通过系谱数据计算两个亚群遗传关联均为 0，表明由于世代分隔较远，两个群体从系谱衡量已没有遗传联系。但基于基因组信息的不同关系矩阵，r 和 CR 结果均不为零，表明基因组数据能够捕捉系谱中不存在的遗传关联。r 和 CR 基于 **G** 和 **H** 很低，分别为(0.0008，0.0003)和(0.003，0.02)，基于 $\mathbf{G_{0.5}}$ 则高达 0.75 和 0.91。r 基于 $\mathbf{G_s}$ 由于出现负值导致不可计算，CR 则与基于 **G** 阵接近。与 r 和 CR 相比，决定系数 CD 所有情况下都较高，在 0.59~0.69 之间，**G** 阵高于 **A** 阵，$\mathbf{G_s}$ 阵最高。对于大多数遗传关联估计方法，**H** 阵结果均介于 **A** 阵与 **G** 阵之间。

表 2 反映了基于 3 家核心场的群体关联估计方法和关系矩阵对群体关联大小的影响。由于 3 个场之间没有系谱联系，没有考虑综合系谱和基因组信息的 H 矩阵。主成分分析表明 3 个群体在基因组信息上存在联系，如图 1 所示，养猪中心与北京六马群体都为美系大白，遗传背景较为接近，新疆天康和养猪中心群体分化最大。场间关联结果也基本表明，大多数情况下养猪中心与北京六马群体关联更高些。在 PEVD、PEVD(x)和 VED 三种方法中，由于没有系谱联系，基于 A 阵的 PEVD 最大，例外情况是，养猪中心与新疆天康之间的遗传关联，基于 **G** 阵和 $\mathbf{G_s}$ 阵计算的 PEVD 高于 **A** 阵。所有情况下，**G** 阵和 $\mathbf{G_s}$ 阵 PEVD 结果接近，基于 $\mathbf{G_{0.5}}$ 的 PEVD 最小。与模拟数据结果类似，PEVD(x)和 VED 方法基于 **G**、$\mathbf{G_s}$、$\mathbf{G_{0.5}}$ 阵结果劣于 **A** 阵，基于 **A** 阵的 PEVD(x)和 VED 为 0.02~0.06，远低于 **G** 阵及其扩展矩阵。在不同 **G** 阵结果中，$\mathbf{G_{0.5}}$ 阵 PEVD(x)和 VED 最小，但对于养猪中心与新疆天康，$\mathbf{G_{0.5}}$ 阵 PEVD(x)和 VED 高于 **G** 阵与 $\mathbf{G_s}$ 阵，所有情况下，**G** 阵与 $\mathbf{G_s}$ 阵结果类似。基于 **A** 阵计算的 3 家核心场之间的预测误差相关 r 和关联率 CR 均为 0，但决定系数 CD 较高，在 0.55~0.67 之间，与模拟数据结果反映的趋势相似。使用基于基因组信息的 **G** 阵及其校正矩阵计算的 r 和 CR 都不为零，3 个场基于 **G** 阵的 r 均为 0.01，CR 分别为 0.15、0.07 和 0.04。3 个场基于 $\mathbf{G_s}$ 的 r 和 CR 与基于 **G** 阵接近，但 3 个场基于 $\mathbf{G_{0.5}}$ 的 r 和 CR 很高，分别为(0.59，0.49，0.48)和(0.94，0.82，0.82)。同时，3 个场基于 **G**、$\mathbf{G_s}$、$\mathbf{G_{0.5}}$ 的 CD 值与基于 **A** 阵相差不大，在 0.59~0.68 之间变化。

2.2 世代对群体关联影响

表 1 中模拟数据结果表明两个亚群经过多个世代分离后，系谱上很难建立群体间遗传联系，但基因组信息仍能捕获到群体间联系。随着世代增加，

表 1 不同群体关联估计方法基于关系矩阵 A、G 和 H 群体遗传关联汇总(模拟数据第 5 世代)

Table 1 Summary of genetic connectedness obtained by different approaches based on relationship matrices A,G and H in the 5th generation of simulated data

方法	关系矩阵				
	A	G	$G_{0.5}$	G_s	H
PEVD	1.6471	1.3218	0.9285	1.5287	1.3725
PEVD(x)	0.0003	0.3248	0.2662	0.4165	0.0016
VED	0.0019	0.3279	0.2690	0.4196	0.0037
CD	0.5882	0.6896	0.6790	0.7366	0.6550
r	0	0.0008	0.7478	NaN	0.0003
CR	0	0.0031	0.9109	0.0035	0.0200

NaN 表示因 r 分母中出现负值导致不可计算，表 4 同。

表 2　3 家猪育种场基于不同估计方法和关系矩阵 A、G 遗传关联汇总

Table 2　Summary of genetic connectedness obtained by different approaches based on relationship matrices A and G between three pig breeding farms

群体	方法	关系矩阵			
		A	G	G_s	$G_{0.5}$
BJLM-BBSC	PEVD	15.5600	11.9300	12.3100	9.0500
	PEVD(x)	0.0200	1.0600	1.1100	0.8600
	VED	0.0500	1.0800	1.1300	0.8800
	CD	0.5500	0.6700	0.6800	0.6700
	r	0	0.0100	0.0200	0.5900
	CR	0	0.1500	0.1500	0.9400
BJLM-XJTC	PEVD	15.1500	14.1300	13.4300	11.3200
	PEVD(x)	0.0200	2.5200	2.3300	2.3000
	VED	0.0400	2.5400	2.3500	2.3300
	CD	0.5600	0.6200	0.6100	0.6200
	r	0	0.0100	0.0100	0.4900
	CR	0	0.0700	0.0700	0.8200
BBSC-XJTC	PEVD	13.2300	14.5100	13.9800	12.200
	PEVD(x)	0.0300	2.1600	2.0500	2.4800
	VED	0.0600	2.1900	2.0800	2.5100
	CD	0.6200	0.6000	0.5900	0.5900
	r	0	0.0100	0.0100	0.4800
	CR	0	0.0400	0.0400	0.8200

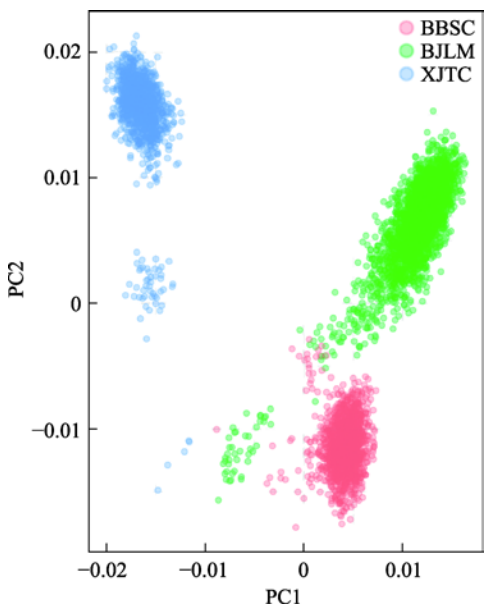

图 1　3 家核心场芯片数据主成分分析结果

Fig. 1　Principal component analysis of three pig breeding farms

PC1、PC2 分别为第一和第二主成分，BBSC、BJLM、XJTC 分别代表北京六马、养猪中心及新疆天康 3 家国家生猪核心育种场。

两个亚群遗传结构差异会越来越大，群体间遗传关联也会减弱。如表 3 所示，随着世代增加，所有群体关联估计方法基于 **G** 阵结果均显示群体间遗传关联程度不断降低。PEVD、PEVD(x)和 VED 估计育种值预测误差，从第 5 世代到第 7 世代，两个亚群 PEVD、PEVD(x)和 VED 逐渐增大，说明预测误差

表 3　不同世代基于 G 阵的遗传关联变化情况

Table 3　The changes of genetic connectedness based on G matrix in different generations of simulated data

方法	世代		
	5	6	7
PEVD	1.3218	1.4136	1.5150
PEVD(x)	0.3248	0.4380	0.5835
VED	0.3279	0.4412	0.5866
CD	0.6896	0.6733	0.6616
r	0.0008	0.0001	0
CR	0.0031	0.0002	0

变大。r 和 CR 则从 5 世代的 0.0008 和 0.0031 逐渐下降至第 7 世代的 0，说明随着群体不断分化，两个群体之间的遗传距离越来越远。与其他方法不同，CD 变化幅度较小，仅从第 5 世代的 0.6896 减小到第 7 世代的 0.6616。

2.3 遗传力对遗传关联统计量的影响

本研究采用模拟数据中第 5 世代数据，通过设置不同的遗传力大小(0.1、0.3、0.5 和 0.7)，研究遗传力大小对遗传关联统计量的影响。如表 4 所示。随着性状遗传力从 0.1 增加到 0.7，除了 PEVD(x)方法基于 **A** 阵不同遗传力下群体关联值保持为 0.0003 不变之外，使用不同关系矩阵 **A**、**G**、和 **H** 计算 PEVD、PEVD(x)和 VED 均不断降低。CD 基于不同关系矩阵的值也是随着遗传力变大而变大，r 和 CR 基于 **A** 阵的值在不同遗传力水平下为 0，它们基于 **G** 阵随遗传力变大而变大，但 CR 基于 **G** 阵变化幅度较小。从表 4 也可以看出，当遗传力为 0.3~0.7 时，在相同遗传力水平下，PEVD 基于 **G** 阵优于 **A** 阵，**H** 阵介于两者之间；对低遗传力 0.1，**A** 阵最优，**G** 阵最差。CD 也呈现与 PEVD 相同的现象。PEVD(x)、VED 则是所有遗传力水平下，基于 **G** 阵的值最大，分别在 0.22~0.4 和 0.22~0.42 之间变动，远远大于基于 **A** 阵和 **H** 阵的 0.0003~0.0095 和 0.0006~0.0063 和 0.001~0.012。

表 4 虽然说明随着遗传力变大，PEVD、PEVD(x)和 VED 减小，CD、r、CR 变大，但不意味着群体关联增强。

表 5 表示不同遗传力水平下群体内 PEVD 变化情况，所有关系矩阵群体内 PEVD 值几乎均随着遗传力的增加而降低。可以看出，群体内 PEVD 变化趋势与群体间 PEVD 一致。群体内个体之间的遗传关联程度远高于群体间个体之间，这表明无论个体间有无实质遗传关联，提高遗传力水平对于 PEVD 均有类似的降低作用。因此，由于高遗传力造成的 PEVD 的降低，不能说明群体间关联程度有提高。

表 4 不同遗传力水平下各群体关联估计方法基于关系矩阵 A、G 和 H 结果汇总

Table 4 Summary of genetic connectedness obtained by different approaches based on relationship matrices A, G and H under different levels of heritability

方法	关系矩阵	遗传力(h^2)			
		0.1	0.3	0.5	0.7
PEVD	A	1.8900	1.6500	1.3300	0.9200
	G	2.1900	1.3200	0.8900	0.3600
	H	1.9000	1.3700	1.0100	0.6500
PEVD(x)	A	0.0003	0.0003	0.0003	0.0003
	G	0.4000	0.3200	0.2700	0.2200
	H	0.0095	0.0016	0.0007	0.0003
VED	A	0.0063	0.0019	0.0010	0.0006
	G	0.4200	0.3300	0.2700	0.2200
	H	0.0120	0.0040	0.0020	0.0010
CD	A	0.5300	0.5900	0.6700	0.7700
	G	0.4900	0.6900	0.7900	0.9200
	H	0.5200	0.6600	0.7500	0.8400
r	A	0	0	0	0
	G	0.0003	0.0008	NaN	NaN
	H	0.0003	0.0003	NaN	NaN
CR	A	0	0	0	0
	G	0.0020	0.0030	0.0030	0.0100
	H	0.1500	0.0200	0.0090	0.0050

表 5　不同遗传力水平下群体内个体关联均值(基于 PEVD)

Table 5　Averaged genetic connectedness (PEVD) within group under different levels of heritability

群体	关系矩阵	遗传力			
		0.1	0.3	0.5	0.7
1	**A**	1.8900	1.6500	1.3300	0.9200
	G	1.7400	0.9600	0.6300	0.0800
	H	1.8800	1.1000	0.6900	0.3700
2	**A**	1.8900	1.6500	1.3300	0.9200
	G	1.8300	1.0400	0.6000	0.2100
	H	1.8900	1.6500	1.3300	0.9200

3　讨论

3.1　高密度 SNP 标记估计群体关联的优势

通过系谱数据估计群体遗传关联程度时，一个常见的问题是系谱不全或存在错误，或者无法从系谱中追溯联系。本研究表明，使用基因组数据能够捕捉系谱中不存在的、由更久远的共同祖先导致的个体间遗传关联。即使根据系谱能够建立群体关联，与基于系谱构建的 **A** 矩阵相比，基因组数据可以更加准确地估计个体间亲缘关系[10]，提高群体关联估计准确性。本研究模拟数据和实际数据结果都显示，大部分遗传关联估计方法基于高密度 SNP 标记建立的个体亲缘关系矩阵都优于基于 **A** 矩阵。这与 Yu 等[16]、Zhang 等[17]研究结果一致，说明利用 SNP 标记估计群体关联更有优势。

3.2　群体间遗传关联度量方法比较

PEVD(x)和 VED 方法为 PEVD 方法的近似估计方法，这两种方法相比于 PEVD 方法计算简单，但本研究模拟数据和实际数据结果表明，相同条件下 PEVD(x)和 VED 均小于 PEVD (表 1,表 2),PEVD(x)和 VED 基于 **G**、$\mathbf{G}_s$、$\mathbf{G}_{0.5}$及 **H** 阵结果劣于 **A** 阵，基于 **A** 阵的 PEVD(x)和 VED 过低,接近于 0 (表 1，表 2)，说明两个群体个体间育种值预测误差很小，这与实际情况有很大偏离。而且，当遗传力从 0.1 提高到 0.7，PEVD(x)方法基于 **A** 阵一直保持为 0.0003，但基于 **G** 阵却在变小(表 4)，说明 PEVD(x)和 VED 不是理想的度量群体关联的方法。

PEVD 及其近似估计方法的一个缺点是取值没有范围，如表 1 和表 2 所示，模拟数据与真实数据估计值差异很大，因此难以判断遗传关联程度。另外 PEVD 容易受到群体大小和结构的影响,例如两个群体基于背膘厚性状计算得到的 PEVD 为 0.8 mm，这个结果对于两个大群体而言可能表示关联程度较差，但是对于两个小群体可能表示关联程度较好[7]。CD、r 和 CR 方法取值范围在 0~1 之间，可以比较好度量群体关联。但是 CD 值即使系谱上不存在遗传联系仍然很高，而 CR 和 r 为 0 (表 1，表 2)。当估计养猪中心和新疆天康群体关联时，CD 基于 **A** 阵最高(表 2)，与其他统计量不太一样，表明 CD 容易高估群体关联程度。统计量 r 大多数情况下低于 CR，但是在实际数据中，不能准确反映群体间的实际群体关联。当用 r 度量养猪中心、北京六马和新疆天康 3 个群体间遗传关联时，基于 **G** 阵的 r 值均为 0.1，区分不出群体的分化远近。而养猪中心–北京六马、北京六马–新疆天康、养猪中心–新疆天康基于 **G** 阵的 CR 分别为 0.15、0.07、0.04，能很好说明群体之间的遗传关联情况。越来越多研究表明，CR 可以作为衡量群体关联程度的稳定方法[21]，MATHUR 等[22]利用加拿大育种数据进行分析，结果显示场间平均遗传关联 CR 大于等于 0.03 时开展联合遗传评估效果较好。这表明虽然通过系谱无法开展 3 个核心场间的联合遗传评估，但是可以开展基于基因组信息的基因组联合评估，如北京地区的大白猪基因组联合育种[23]。而且与 PEVD 相比，CR 不需要进行个体间两两匹配求均值，计算简单，并且可以同时估计多个群体之间的遗传关联程度。

3.3　基于 SNP 标记关系矩阵比较

基于 SNP 标记构建的个体关系矩阵可以更真实反映个体间亲缘关系，但是要求每个标记的等位基

因频率为基础群体的，这个不易获得，所以通常用当前群体的等位基因频率代替。因此除了经典的 **G** 阵，还有其他方法来解决等位基因频率问题，如 $\mathbf{G}_s$ 和 $\mathbf{G}_{0.5}$。本研究表明，基于 **G** 阵与 $\mathbf{G}_s$ 阵的各种遗传关联估计方法结果比较接近，$\mathbf{G}_{0.5}$ 过于高估群体间遗传关联。当用 CR 度量模拟数据两个亚群和实际数据 3 个核心场间遗传关联时，基于 **A** 阵的群体关联都为 0，说明群体间联系很弱，基于 **G** 阵与 $\mathbf{G}_s$ 的模拟数据亚群分别为 0.0031 和 0.0035，但基于 $\mathbf{G}_{0.5}$ 则高达 0.91；3 个核心场间基于 **G** 阵与 $\mathbf{G}_s$ 均为 0.15、0.07 和 0.04，而基于 $\mathbf{G}_{0.5}$ 则为 0.94、0.82 和 0.82。$\mathbf{G}_{0.5}$ 阵假定所有标记的最小等位基因频率均为 0.5，此假设过于理想，既无法反映基础群体的情况，也无法反映当前群体的真实情况，从而导致遗传关联结果出现较大偏差。因此，$\mathbf{G}_{0.5}$ 阵不适合用于评估群体遗传关联。

本研究中 **H** 阵结果一般介于 **A** 阵和 **G** 阵之间，这与 Yu 等研究结果相同[16]。**H** 阵由 **A** 阵和 **G** 阵混合而成，因此使用 **H** 阵估计遗传关联结果一般优于仅使用系谱数据结果，而当大部分个体均有基因组数据时，**H** 阵遗传关联结果提升幅度可能低于 **G** 阵。

3.4 遗传力影响

本研究设定了高、中、低 4 种遗传力水平检验其对群体关联估计方法影响。大多数情况下，群体关联统计量会随着遗传力升高而改善，但就像群体内个体遗传关联也呈现相同变化一样(表 5)，不能说明群体间关联程度有提高。遗传力升高会提高育种值估计准确性，降低了育种值预测误差，因而改善了相应的群体关联统计量。因此，在育种实践中，遗传力不同的性状估计的遗传关联结果之间缺乏可比性。另外，本研究发现，低遗传力(0.1)情况下，基于 **A** 阵的 PEVD 优于 **G** 阵，与大多数情况下 **G** 阵优于 **A** 阵相反，说明低遗传力性状不太适合用来估计群体遗传关联。

参考文献(References):

[1] Mathur PK, Sullivan BP, Chesnais JP. Measuring connectedness: concept and application to a large industry breeding program. *Proc 7th World Congr Genet Appl to Livest Prod*, 2002, 19: 23. [DOI]

[2] Zhang JX, Zhang SY, Qiu XT, Gao H, Wang CC, Wang Y, Zhang Q, Wang ZG, Yang HJ, Ding XD. The genetic connectedness of duroc, landrace and yorkshire pigs in China. *Acta Vet Et Zootech Sin*, 2017, 48(9): 1591–1601.
张金鑫，张锁宇，邱小田，高虹，王长存，王源，张勤，王志刚，杨红杰，丁向东. 我国杜洛克、长白和大白猪场间遗传联系分析. 畜牧兽医学报，2017，48(9)：1591–1601. [DOI]

[3] Gao H, Qiu XT, Wang CC, Zhang JX, Zhang SY, Wang Y, Zhang Q, Wang ZG, Yang HJ, Ding XD. The regional joint genetic evaluation of duroc, landrace and yorkshire pigs in China. *Acta Vet Et Zootech Sin*, 2018, 49(12): 2567–2575.
高虹，邱小田，王长存，张金鑫，张锁宇，王源，张勤，王志刚，杨红杰，丁向东. 我国杜洛克、长白、大白猪区域性联合遗传评估研究. 畜牧兽医学报，2018，49(12)：2567–2575. [DOI]

[4] Kennedy BW, Trus D. Considerations on genetic connectedness between management units under an animal model. *J Anim Sci*, 1993, 71(9): 2341. [DOI]

[5] Laloë D. Precision and information in linear models of genetic evaluation. *Genet Sel Evol*, 1993, 25(6): 557–576. [DOI]

[6] Lewis RM, Crump RE, Simm G,Thompson R. Assessing connectedness in across-flock genetic evaluations. In: Proceedings of the British Society of Animal Science. Scarborough, 22–24 March, 1999, 121–122. [DOI]

[7] Mathur PK, Sullivan B, Chesnais J. Estimation of the degree of connectedness between herds or management groups in the canadian swine population. 2002. [DOI]

[8] Wang AG, Laloe D, Schaeffer LR. Measures of genetic connectedness between herds in swine under mixed linear models. *Hereditas (Beijing)*, 2000, 22(5): 295–297.
王爱国，Laloe D.，Schaeffer LR. 混合线性模型下猪群间遗传联系的度量. 遗传，2000，22(5)：295–297. [DOI]

[9] Muir WM. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *J Anim Breed Genet*, 2007, 124(6): 342–355. [DOI]

[10] Daetwyler HD, Villanueva B, Bijma P, Woolliams JA. Inbreeding in genome-wide selection. *J Anim Breed Genet*, 2007, 124(6): 369–376. [DOI]

[11] Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF. Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, 2008, 178(1): 553–561. [DOI]

[12] Zhang Z, Li X, Ding X, Li J, Zhang Q. GPOPSIM: a

simulation tool for whole-genome genetic data. *BMC Genet*, 2015, 16(1): 1–6. [DOI]

[13] Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*, 2009, 84(2): 210–223. [DOI]

[14] Vanraden PM. Efficient methods to compute genomic predictions. *J Dairy Sci*, 2008, 91(11): 4414–4423. [DOI]

[15] Fernando RL, Cheng H, Garrick DJ. An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. *Genet Sel Evol*, 2016, 48(1): 80. [DOI]

[16] Yu H, Spangler ML, Lewis RM. Genomic relatedness strengthens genetic connectedness across management units. *G3-Genes Genom Genet*, 2017, 7(10): 3543–3556. [DOI]

[17] Zhang SY, Olasege BS, Liu DY, Wang QS, Pan YC, Ma PP. The genetic connectedness calculated from genomic information and its effect on the accuracy of genomic prediction. *PLoS One*, 2018, 13(7): e0201400. [DOI]

[18] Toro MA, García-Cortés LA, Legarra A. A note on the rationale for estimating genealogical coancestry from molecular markers. *Genet Sel Evol*, 2011, 43(1): 1–10. [DOI]

[19] Vitezica ZG, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations under selection. *Genet Res*, 2011, 93(5): 357–366. [DOI]

[20] Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. *J Dairy Sci*, 2009, 92(9): 4656–4663. [DOI]

[21] Zhang Q, Ding XD, Chen YS. Development and Application of Swine Genetic Evaluation System in China. *Chin J Anim Sci*, 2015, 51(08): 61–65.
张勤, 丁向东, 陈瑶生. 种猪遗传评估技术研发与评估系统应用. 中国畜牧杂志, 2015, 51(08): 61–65. [DOI]

[22] Mathur PK, Ssllivan BP, Chesnais JP. Measuring connectedness: concept and application to a large industry breeding program. In: Proceedings of 7th world congress on genetics applied to livestock production. Montpellier, 19–23 August, 2002. [DOI]

[23] Zhang JX, Tang SQ, Song HL, Gao H, Jiang Y, Jiang YF, Mi SR, Meng QL, Yu F, Xiao W, Yun P, Zhang Q, Ding XD. Joint genomic selection of Yorkshire in Beijing. *Sci Agric Sin*, 2019, 52(12): 2161–2170.
张金鑫, 唐韶青, 宋海亮, 高虹, 蒋尧, 江一凡, 弥世荣, 孟庆利, 于凡, 肖炜, 云鹏, 张勤, 丁向东. 北京地区大白猪基因组联合育种研究. 中国农业科学, 2019, 52(12): 2161–2170. [DOI]

(责任编委: 李明洲)