



基于单倍型匹配率鉴定系谱外祖父的准确性研究

娄建东¹, 司敬方¹, 田佳², 侯丽娥², 周佳敏², 王雅春¹, 张毅^{1*}

(1. 中国农业大学动物科学技术学院, 北京 100193; 2. 宁夏回族自治区畜牧工作站, 银川 750105)

摘要: 旨在评估当母亲基因型数据缺失的情况下, 不同芯片密度和单倍型窗口长度对外祖父识别准确性的影响, 为利用基因组信息准确构建系谱关系提供可靠技术手段。本研究基于 6 117 头奶牛的基因组芯片数据, 分别构建低密度芯片(5K、10K)和中等密度芯片(20K、50K)数据集, 并设置 5、10、35、75 和 120 五种 SNP 窗口长度, 将全基因组划分为连续的单倍型片段并进行单倍型推断。随后, 结合子代个体已知的父本基因型信息, 剔除与父源完全一致的单倍型片段以推断母源单倍型序列, 并与候选公牛单倍型逐片段比对, 计算单倍型匹配率, 并选取匹配率最高的个体作为候选外祖父。同时, 选取 345 头谱系信息完整可靠的个体用于方法准确性验证。此外, 研究还进一步计算了个体的亲缘系数与近交系数, 以揭示亲缘关系与近交对鉴定方法的潜在影响。结果表明, 基于单倍型分段匹配的识别方法在绝大多数芯片密度与窗口长度的条件下, 均可实现 90% 以上的鉴定准确率, 其中 20K×35 SNPs 的组合表现最优, 准确率高达 98%, 方法整体表现出较高的准确性与稳定性。然而, 不同组合条件下的最优判定阈值和准确性存在明显差异, 提示芯片密度与窗口长度对方法性能具有重要影响。进一步分析显示, 亲缘关系与近交水平同样显著影响匹配效果与准确性, 提示在实际应用中需综合考量群体遗传背景的影响。综上所述, 该方法在母系信息缺失及大规模实际生产群体背景下展现出较高的鉴定准确性, 可作为外祖父识别的有效手段。此外, 结合群体遗传背景合理选择芯片密度与窗口长度是提升鉴定效果的关键。

关键词: 奶牛; 外祖父; 单倍型匹配; SNP 芯片密度; 窗口长度; 阈值

中图分类号: S823.2

文献标志码: A

文章编号: 0366-6964(2026)02-0716-12

Accuracy of Maternal Grandsire Discovery Based on Haplotype Matching Rate

LOU Jiandong¹, SI Jingfang¹, TIAN Jia², HOU Li'e², ZHOU Jiamin², WANG Yachun¹, ZHANG Yi^{1*}

(1. College of Animal Science and Technology, China Agricultural University, Beijing 100193, China;

2. Ningxia Hui Autonomous Region Animal Husbandry Workstation, Yinchuan 750105, China)

Abstract: This study aimed to develop a haplotype-segment matching-based maternal grandsire discovery approach applicable to different SNP array densities and segment window sizes in the absence of dam genotypes, thereby providing a reliable technical approach for pedigree reconstruction. The study utilized genomic data from 6 117 dairy cows to construct datasets of varying SNP densities, including low-density (5K, 10K) and medium-density (20K, 50K) SNP panels. For each dataset, the genome was partitioned into fixed-length haplotype segments of 5, 10, 35, 75, and 120 SNPs, followed by haplotype inference. Based on the known genotypes of the offspring's sire, paternal haplotypes were excluded to infer maternal haplotypes, which were then compared with those of candidate bulls on a segment-by-segment basis to calculate the haplotype matching rate (HMR). The

收稿日期: 2025-05-26

基金项目: 宁夏回族自治区重点研发项目(2023BCF01004); 宁夏回族自治区农业育种专项(2019NYYZ05); 财政部和农业农村部: 国家现代农业产业技术体系(CARS-36)

作者简介: 娄建东, 硕士生, 主要从事奶牛育种方向的研究, E-mail: 2329421373@qq.com

* 通信作者: 张毅, 主要从事分子数量遗传学研究, E-mail: yizhang@cau.edu.cn

bull with the highest HMR was designated as the putative maternal grandsire (MGS). A total of 345 individuals with complete and validated pedigree information were used to verify method accuracy. Furthermore, genomic relatedness coefficients and homozygosity levels were calculated to investigate the potential effects of kinship structure and inbreeding on identification outcomes. The haplotype-segment matching method achieved more than 90% assignment accuracy across most combinations of SNP-array density and segment window length. The 20K array with a 35-SNPs window yielded the highest accuracy at 98%, demonstrating robust performance. Nonetheless, the optimal decision thresholds for the haplotype match rate and the resulting accuracies varied substantially with parameter settings, indicating that SNP density and window length were key determinants of performance. Additional analyses showed that kinship structure and inbreeding levels significantly affected matching performance and assignment accuracy, suggesting that the population genetic background should be considered in practical applications. The haplotype-segment matching (HSM)-based method demonstrated high accuracy in maternal grandsire identification under conditions of missing maternal genotype data, making it a valuable tool for pedigree correction in large-scale commercial herds. Optimizing SNP density and haplotype window length based on the genetic background of the population is essential to enhance identification efficiency and robustness in practical applications.

Keywords: dairy cattle; maternal grandsire; haplotype matching; SNP chip density; window length; threshold

* **Corresponding author:** ZHANG Yi, E-mail: yizhang@cau.edu.cn

近年来,随着基因组技术的快速发展与大规模遗传评估体系的完善,大量公牛的基因分型数据得以迅速积累,使基于单核苷酸多态(single nucleotide polymorphism, SNP)纯合位点冲突的常规亲子鉴定方法得以广泛应用^[1-4]。目前,该方法已帮助数以百万计的公牛父亲身份得到准确校正,为全球奶牛的遗传评估奠定了可靠父系基础^[5-6]。然而,由于检测成本等因素的考量,遗传价值相对较低的母牛个体未能实现大规模的基因分型检测,导致在实际生产中母系谱系仍存有大量的缺失和错误^[7-9]。截至2018年,美国奶牛育种委员会(Council on Dairy Cattle Breeding, CDCB)数据库记录的约220万头已基因分型奶牛个体中,父亲身份验证率高达97%,而母亲身份验证率仅为39%^[10-11]。因此,大量具有重要信息价值的母系祖先身份无法得到及时、准确地校正,在一定程度上限制了母系信息在奶牛遗传评估与育种决策中的有效利用^[12-13]。在此背景下,外祖父(maternal grandsire, MGS)的准确鉴定不仅成为弥补母系信息缺失的关键,还能有效提高近交系数与育种值估计的可靠性^[14]。

目前,SNP位点冲突计数法已用于奶牛外祖父的鉴定,通过统计候选外祖父与待测个体之间的相

反纯合位点数来判断是否存在祖孙关系。然而,基于SNP位点冲突的外祖父鉴定方法需要足够数量的相反纯合子信息,当使用低密度芯片或杂交群体数据时,由于待测SNP总数有限且仅利用纯合位点信息,导致鉴定准确性通常较低^[15-18]。近年来,基于单倍型匹配的外祖父鉴定策略为解决上述问题提供了新思路^[19]。该方法的核心步骤如下:1)首先,对全部个体采用固定窗口分段的策略将基因组划分为连续的单倍型片段;2)随后,结合个体已知的父本基因型信息,剔除与父源单倍型完全一致的片段,从而推断获得高分辨率的母源单倍型序列;3)随后,将待测个体的母源单倍型与候选公牛的单倍型逐段比对,并计算单倍型匹配率;4)当候选个体与待测个体之间的单倍型匹配率(haplotype match rate, HMR)超过预设的经验阈值(如 $\geq 35\%$)时,即判定该候选公牛为其潜在的外祖父。与传统方法相比,基于单倍型推断的策略能有效利用包括杂合位点在内的全部基因型信息,从而显著提高鉴定的准确率;此外,该方法可在已知的外祖父基础上进一步扩展至外曾祖父(maternal great-grandsires, MGGS)乃至更远祖先的鉴定。Nani等^[14]已将该方法整合至美国CDCB的日常分析流程中,为29万头

母系信息缺失的个体成功鉴定了外祖父,并通过“虚拟母本 ID”实现谱系自动链接,使传统估计育种值(estimated breeding value, EBV)可靠性提高了约 6 个百分点。尽管单倍型方法已在实际应用中展现出巨大潜力,但其准确性仍受多种因素影响,相关机制尚缺乏系统研究,主要包括:1)目前尚缺乏芯片密度与单倍型窗口长度对鉴定准确率影响的系统性研究;2)外祖父鉴定的判定阈值多依赖经验,缺乏对于不同信息背景下的最优阈值界定依据;3)群体内的亲缘结构及近交水平对单倍型方法鉴定的影响尚缺乏研究分析。

针对上述问题,本研究开展以下工作:1)系统评估芯片密度与单倍型窗口长度对外祖父鉴定准确性的影响;2)通过真实数据验证,明确不同信息条件下的最优阈值范围,并系统评估信息差异对阈值设定的影响;3)分析群体亲缘结构及近交水平对单倍型方法鉴定结果的潜在影响,为不同遗传背景

下的应用提供参考。旨在系统评估单倍型法在不同信息条件下外祖父鉴定的准确性与可行性,为构建低成本高效的外祖父鉴定方法提供参考。

1 材料与方法

1.1 试验群体的构建

本研究通过严格的系谱记录与基因组亲子鉴定筛选出 345 头系谱信息真实、完整的荷斯坦母牛作为待测子代群体,并进一步划分为:1)阳性验证组,包括待检验的子代个体及其对应的真实外祖父,共计 224 对“子代-外祖父”祖孙关系;2)阴性对照组,仅包含 121 头真实外祖父不在祖先候选群中的子代个体,以评估方法在真实外祖父缺失情况下的误识风险,即假阳性率(表 1)。待测子代群体的真实父亲(164 头)、真实外祖父(98 头)和其他无关公牛(5 510 头),合计共 5 772 头荷斯坦公牛组成候选的父亲和外祖父群体(表 1)。

表 1 试验群体构成与样本数量

Table 1 Composition and sample size of the experimental population

组别 Group	待测子代 Offspring	对应父本 Sires of offspring	真实外祖父 Confirmed MGS
阳性验证组 Positive group	224	101	98
阴性对照组 Negative group	121	63	0
合计 Total	345	164	98

1.2 芯片分型位点的挑选与质控

1.2.1 芯片分型位点的挑选 本研究所有样本均通过基因芯片完成分型,所采用的芯片包括纽勤(Neogen) 150K、因美纳(Illumina) 50K 和硕腾(Zoetis) 50K。为统一分析标准,提取所有芯片共有的 20K SNP(19 277 个位点)作为基础集合,依据染色体长度比例和等间距抽样策略构建 5K(5 479 个位点)与 10K(10 255 个位点)低密度子集(公式 1、2);同时,以 Illumina 50K_v2 作为参考模板,并使用 Beagle 软件对 20K 的芯片数据集进行填充,最终生成包含 49 768 位点的 50K 中等密度数据集。

$$n_{(selected)}^{(c)} = \left[\left(n_{(chr)}^{(c)} \right) / \left(\sum_c n_{(chr)}^{(c)} \right) \right] \times N_{total} \quad (\text{公式 1})$$

其中, $n_{(chr)}^{(c)}$ 为染色体 c 的原始 SNP 数目, $\sum_c n_{(chr)}^{(c)}$

为全基因组 SNP 总数, N_{total} 为目标 SNP 总量(5 000 或 10 000), $n_{(selected)}^{(c)}$ 为从每个染色体计划抽取的

SNPs 个数。

$$P_i = P_{min} + i(P_{max} - P_{min}) / (n_{(selected)}^{(c)} + 1), \\ i = 1, 2, \dots, n_{(selected)}^{(c)} \quad (\text{公式 2})$$

其中, P_{min} 为该染色体所有 SNP 的起始位点, P_{max} 为终止位点, $n_{(selected)}^{(c)}$ 为根据式(1)得到的需在染色体 c 上保留的 SNP 数量, P_i 为计算得到的抽取位点坐标。

1.2.2 基因组数据质控 所有数据集均使用 PLINK 软件进行质量控制,参数设定为: --maf 0.01, --geno 0.01, --hwe 1e-6。质控后,所有个体均通过筛选,各基因分型数据集分别保留 4 942(5K)、9 411(10K)、18 497(20K)和 43 786(50K)个位点用于后续单倍型推断与外祖父鉴定分析。

1.3 固定窗口长度下的单倍型分段与构建

为系统评估单倍型窗口长度对外祖父鉴定的影响,本研究采用美国农业研究局(USDA-ARS)研发的 Findhap(v3)程序,通过固定窗口分段策略进行单倍型推断。该程序将全基因组划分为连续的

单倍型片段,并结合群体聚类与系谱等先验信息进行迭代优化,以获得分段后的单倍型结果^[20-21]。随后,在4种芯片密度(5K、10K、20K、50K)数据集上,分别设置5、10、35、75和120个SNPs五种窗口长度,进行单倍型推断并用于后续外祖父鉴定分析。

1.4 外祖父鉴定方法与准确性评估

本研究基于分段单倍型比对方法,构建候选外祖父鉴定流程。首先,对每个个体的全基因组按照预设的固定窗口长度(5、10、35、75和120个SNPs)进行分段,构建单倍型片段序列。随后,结合目标子代个体已知的父本基因型信息,通过剔除与父本完全一致的单倍型片段,推断获得其母源单倍型序列。以该母源单倍型为参考模板,在候选公牛群体中逐一比对,并设置候选外祖父的年龄限制,即候选个体的出生日期须早于子代个体至少3年,以避免同胞等近亲个体对鉴定结果的干扰。最后统计各候选个体与子代之间在全基因组范围内匹配的单倍型片段数,并定义单倍型匹配率(HMR)如下:

$$HMR = HAP_m / HAP_t \quad (\text{公式3})$$

其中, HAP_m 为单倍型匹配数, HAP_t 为全基因组的单倍型分段总数。对于每个子代个体,在所有候选个体中选取HMR值最高者,作为其潜在的候选外祖父。

为确定该方法的最优判定标准,本研究在观察到的HMR值范围内,以0.1为间隔递增设定HMR阈值 t ,并在每个阈值下进行分类判定:当 $HMR \geq t$ 时,判定存在外祖父关系,反之则判定为非外祖父。在每一阈值设定下,依据个体与真实外祖父的关系,将鉴定结果划分为以下4类:真阳性(true positive, TP),即真实存在外祖父关系且被正确识别;假阳性(false positive, FP),即实际不存在外祖父关系但被错误识别为存在;真阴性(true negative, TN),即真实不存在外祖父关系且未被错误识别;假阴性(false negative, FN),即真实存在外祖父关系但未能被正确识别。在此分类基础上,计算各阈值下的鉴定准确率(accuracy)作为方法性能指标:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (\text{公式4})$$

通过遍历全部阈值,确定准确率达到最大时所对应的HMR阈值,并据此设定最优阈值判定界限。

1.5 基因组亲缘系数与近交系数

为评估亲缘关系与近交水平对单倍型鉴定结果的影响,本研究基于质控后的20K SNP数据集(共18 497个位点),使用PLINK v1.9的--genome命令计算个体间的IBD(identity-by-descent)共享概率,以衡量其基因组亲缘关系^[22-24]。采用--het命令计算个体杂合度,并以其互补值表示纯合度(homozygosity),用于评估个体的基因组近交水平^[25-26]。

为进一步比较不同分类结果下个体的亲缘关系与近交水平差异,依据设定的HMR阈值将所有“个体-外祖父”组合划分为真阳性(TP)、假阳性(FP)、假阴性(FN)和真阴性(TN)4类,分别提取各类样本并计算其与候选外祖父之间的亲缘系数与纯合度水平。

2 结 果

2.1 单倍型窗口长度与芯片密度对HMR分布的影响

HMR的分布特征在一定程度上反映了方法的判别能力,并揭示芯片密度、窗口长度及遗传背景等因素对鉴定效果的潜在影响。为此,本研究以345头系谱完整的待测子代为研究对象,基于推断的母源单倍型与候选公牛逐一比对,选取匹配率最高的个体作为候选外祖父,并根据真实外祖父是否存在于候选群体,将对应的“个体-候选外祖父”组合划分为阳性验证组($n=224$)与阴性对照组($n=121$)。随后,在4种芯片密度(5K、10K、20K、50K)与5种窗口长度(5、10、35、75、120个SNPs)的20种组合下,比较两组HMR的分布特征(图1)。

2.1.1 HMR分布的总体特征 HMR的总体分布特征显示,在不同芯片密度与窗口长度的组合下,除5K和10K芯片在75与120个SNPs的较长窗口下出现明显重叠外,其他组合均呈现出HMR分布峰值的清晰分离,且阳性组中个体的HMR显著高于阴性组,表明在大多数组合下候选外祖父是否在群对应着两个显著不同的匹配率分布模式。此外,在所有组合中均存在不同程度的分布重叠,表明基于匹配率阈值的鉴定方法不可避免地会产生一定比例的假阳性与假阴性个体。

值得注意的是,即便在阳性验证组内部,不同个体间的匹配率水平亦存在显著差异,尤其在10K、

20K 及 50K 芯片密度与窗口长度为 35、75 和 120 个 SNPs 的组合中表现最为明显,部分个体显示出更高的 HMR 值,表明其与外祖父间可能存在更近的亲缘关系以及更高的近交水平。

2.1.2 芯片密度对 HMR 分布特征的影响 当芯片密度由 5K 提升至 20K 时,在中长窗口条件(35、75 和 120 个 SNPs)下,HMR 分布整体向右偏移,阳性组与阴性组之间的重叠区域明显减少,且阳性组中部分个体呈现出匹配率更高的集中分布。表明

芯片密度的增加提高了个体与外祖父之间的 HMR,且两组之间的区分度得到明显提升。相比之下,在短窗口条件(5 和 10 个 SNPs)下,整体分布形态基本保持一致,仅表现为分布的向右偏移,表明短窗口下芯片密度提升对分布的影响较为有限;当芯片密度进一步提高至 50K 时,所有窗口条件下的 HMR 分布形态变化趋于稳定,与 20K 水平基本一致,仅表现为分布的继续右移,表明在芯片密度达到 20K 以上后继续提升密度对 HMR 分布的影响有限。

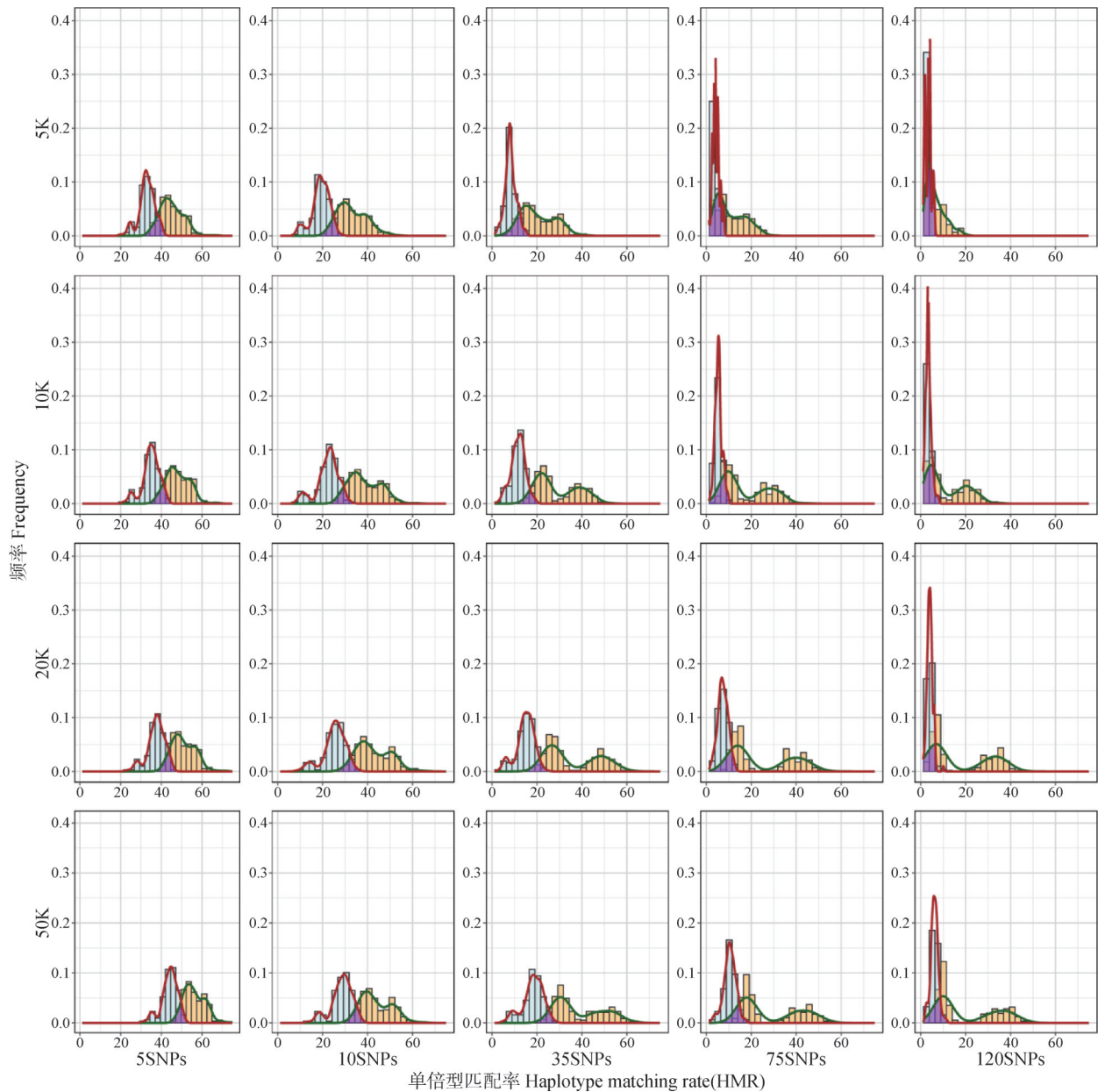


图1 不同芯片密度与窗口长度组合下阳性验证组(黄色)与阴性对照组(蓝色)的单倍型匹配率(HMR)分布图

Fig.1 Distribution of haplotype matching rate (HMR) between the positive validation group (yellow) and the negative control group (blue) under different combinations of SNP chip density and haplotype window length

2.1.3 单倍型分段窗口大小对HMR分布特征的影响 随着窗口长度由5个SNPs增加至35个SNPs,HMR整体分布发生左移,阳性组中个体开始呈现双峰分布,但两组间的分布交汇现象无明显变化,表明在中短窗口条件下,窗口长度的增加虽降低了个体与外祖父之间的HMR,但对组间区分度影响有限。进一步将窗口延长至120个SNPs后,HMR分布继续左移,阳性组与阴性组的分布趋于集中,交汇现象加剧。表明窗口长度的增加进一步降低了个体与外祖父之间的HMR,同时削弱了阳

性组与阴性组个体在HMR上的区分性,使得以固定阈值划分判定结果中的假阳性与假阴性比例的同时上升。

2.2 单倍型窗口长度与芯片密度对鉴定准确率的影响

本研究为分析不同芯片密度与窗口分段长度下的最优鉴定阈值及其对应的最大准确性,在4种芯片密度(5K、10K、20K、50K)与5种窗口长度(5、10、35、75、120个SNPs)的20个组合下分别遍历整个阈值范围并计算其对应的准确性(图2)。

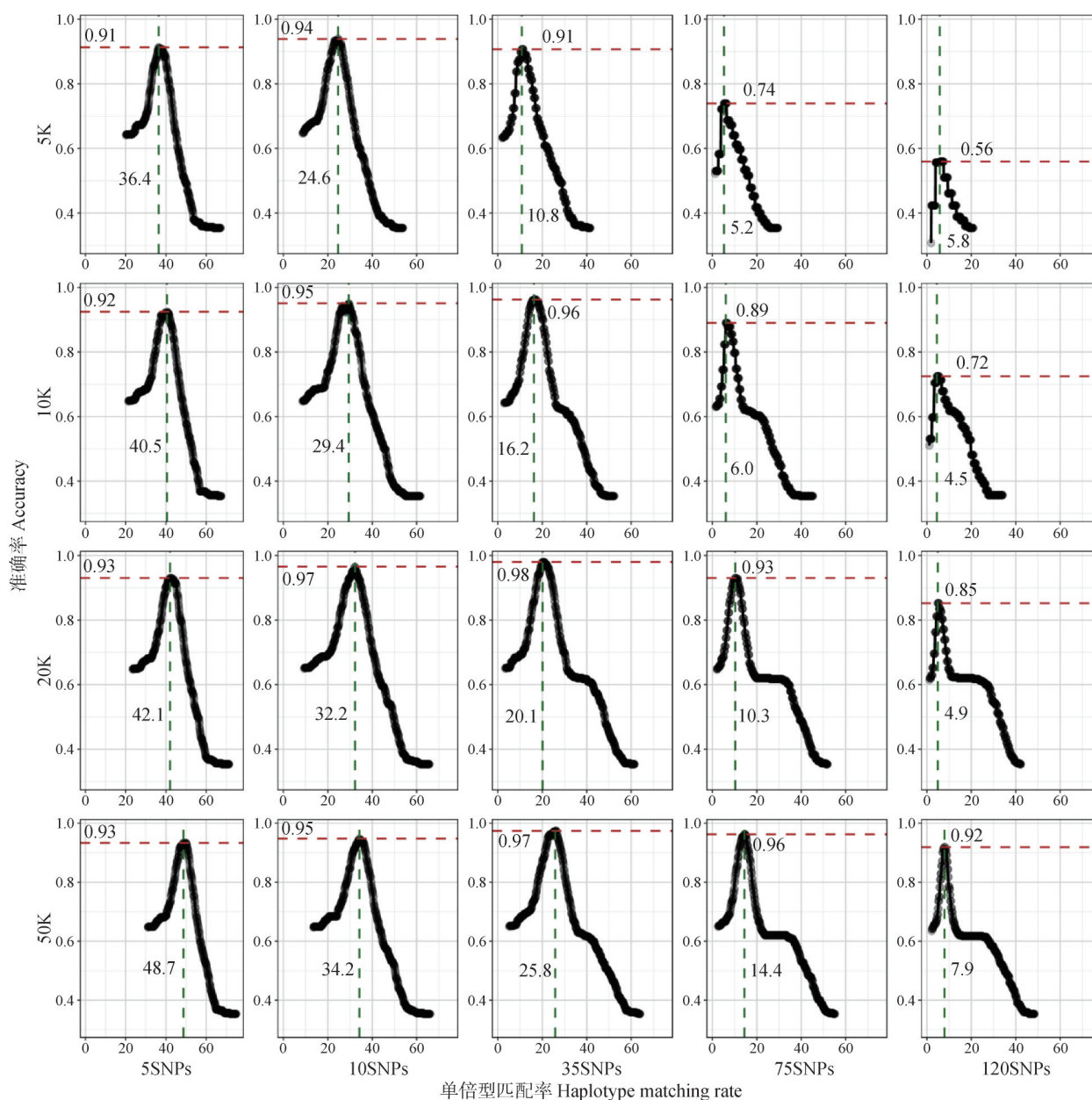


图2 不同芯片密度与窗口长度组合下单倍型匹配率(HMR)与外祖父鉴定准确性的变化曲线

Fig.2 The variation curve of haplotype matching rate (HMR) and maternal grandsire identification accuracy under different combinations of SNP chip density and haplotype window length

2.2.1 鉴定准确率的总体变化趋势 在所有芯片密度与窗口长度组合中,准确率均随匹配率阈值的增加先升高后下降,呈现出单峰曲线的变化特点,表明在各个组合下均存在最优的判定阈值。其中20K芯片密度与35个SNPs窗口长度的组合表现最佳,最高准确率高达98%。同时,不同芯片密度与窗口长度组合下的最优匹配率阈值存在差异,提示在实际应用中需根据芯片信息量与分析策略灵活设定阈值,以获得更高的鉴定准确性和应用效果。

2.2.2 芯片密度对鉴定准确率的影响 随着芯片密度由5K提升至20K,各分段窗口长度下的最大准确率均有所提高,但提升幅度因分段长度而异。在短窗口(5、10个SNPs)下,最优阈值对应的最大准确率提升有限,而在中长窗口(35、75、120个SNPs)下,准确率的提升则更为显著。当芯片密度进一步提高至50K时,75和120个SNPs窗口的最大准确率继续提高,而其余窗口的准确率则略有下降。表明芯片密度对准确率的提升效果在不同窗口长度下存在显著差异。此外,随着芯片密度的增加,达到最大准确率所对应的HMR阈值也随之提高。

2.2.3 单倍型窗口长度对鉴定准确率的影响 在相同芯片密度条件下,最优阈值下的最大准确率随窗口长度的增加呈现出先升高而后下降的趋势特征。在5K芯片密度下,当窗口长度为10个SNPs时最优阈值所对应的准确率达到最高,而在其他芯片密度下,最高准确率均出现在长度为35个SNPs窗口处。这一结果表明,当分段片段过短或过长时,均会导致鉴定准确率下降。此外,随着单倍型窗口长度的增加,达到最大准确率所对应的HMR阈值随之降低。

2.3 群体亲缘结构与近交水平对鉴定效果的影响

2.3.1 群体亲缘结构对鉴定效果的影响 为进一步探究亲缘关系对单倍型鉴定方法的潜在影响,并确保分析结果的稳定性与代表性,本研究采用多阈值划分策略,以避免单一阈值设定下遗漏因HMR阈值波动而易被误判的边界个体。以识别效果最优的20K×35个SNPs组合为例,选取其准确率超过95%时对应的HMR阈值区间(18.8~23.0),并基于区间内的每个阈值将所有“个体-外祖父”组合划分为真阳性(TP)、假阳性(FP)、假阴性(FN)和真阴性(TN)4类。挑选在所有阈值下均被

一致判定为TP(210)或TN(104)的组合,用于代表方法在高置信度条件下可准确识别的稳定组合。FP(17)和FN(14)组则整合区间内所有的误判组合,去重后保留唯一结果,以代表该区间内可能出现的全部错判情况。最终,分别计算4类样本中子代与候选外祖父的基因组亲缘系数(图3)。

Kruskal-Wallis检验结果显示,4组之间存在极显著整体差异($P=1.33\times 10^{-45}$)。其中TP组(0.28 ± 0.06)中的亲缘系数显著高于FN(0.20 ± 0.03 , $P=2.00\times 10^{-4}$)、FP(0.16 ± 0.03 , $P=8.77\times 10^{-8}$)和TN组(0.11 ± 0.07 , $P=3.26\times 10^{-45}$),表明TP组中个体与真实外祖父间具有更近的亲缘关系。FN组的亲缘关系显著低于TP组但仍显著高于TN组($P=0.02$),进一步分析推断发现,FN组中所有个体最高匹配率所对应的候选外祖父均为其真实外祖父,但由于匹配率未达到阈值而被判为阴性,说明FN组中的真实祖孙关系因亲缘水平偏低而被遗漏误判。FP与FN($P=0.37$)、TN($P=0.19$)组的亲缘水平相近,但整体分布仍略高于TN组,表明当个体与候选非真实外祖父之间的亲缘关系较高时可能会导致误判风险的增加。综合FN和FP组的结果,亲缘关系在0.16~0.21范围内的个体组合更易产生假阳性或假阴性的错判。尽管TN组中整体的亲缘系数较低,但仍有少数个体对的亲缘系数过高,进一步分析发现其对应的HMR依然较低,并未被误判为FP,提示单倍型匹配率除亲缘关系外,可能还受其他因素影响。

2.3.2 近交水平对鉴定效果的影响 为进一步分析个体近交水平对单倍型鉴定方法的潜在影响,本研究同样基于20K×35个SNPs组合,在准确率超过95%的HMR阈值区间(18.8~23.0)内,依据前述分类方法划分并选取TP、FP、FN和TN四类个体,计算其纯合度指标(HOM)以评估基因组近交水平(图4)。

Kruskal-Wallis检验结果显示,4组之间同样具有极显著的整体差异($P=1.7\times 10^{-7}$)。其中,TP(0.64 ± 0.02)与FP(0.64 ± 0.01)组的近交水平显著高于TN(0.62 ± 0.05 , $P_{TP-TN}=7.86\times 10^{-8}$, $P_{FP-TN}=0.026$)与FN组(0.62 ± 0.01 , $P_{TP-FN}=0.003$, $P_{FP-FN}=0.035$),表明高近交水平个体的基因组中同源片段比例较高,单倍型一致性增强,匹配率随之升高,因而更易被判定为阳性;而低近交水平个体中重组事件更频繁,单倍型一致性与匹配率较低,更易被判

定为阴性。进一步比较显示,TP与FP之间的近交水平差异不显著($P=0.81$),TN与FN之间亦无显著差异($P=0.54$),说明FP个体在近交水平与匹配率上与TP相近,难以在实际判定中有效区分,因而

更易被误判为假阳性;FN个体与TN接近,同样具有较低的近交水平和匹配率,在固定阈值下更易被误判为假阴性。

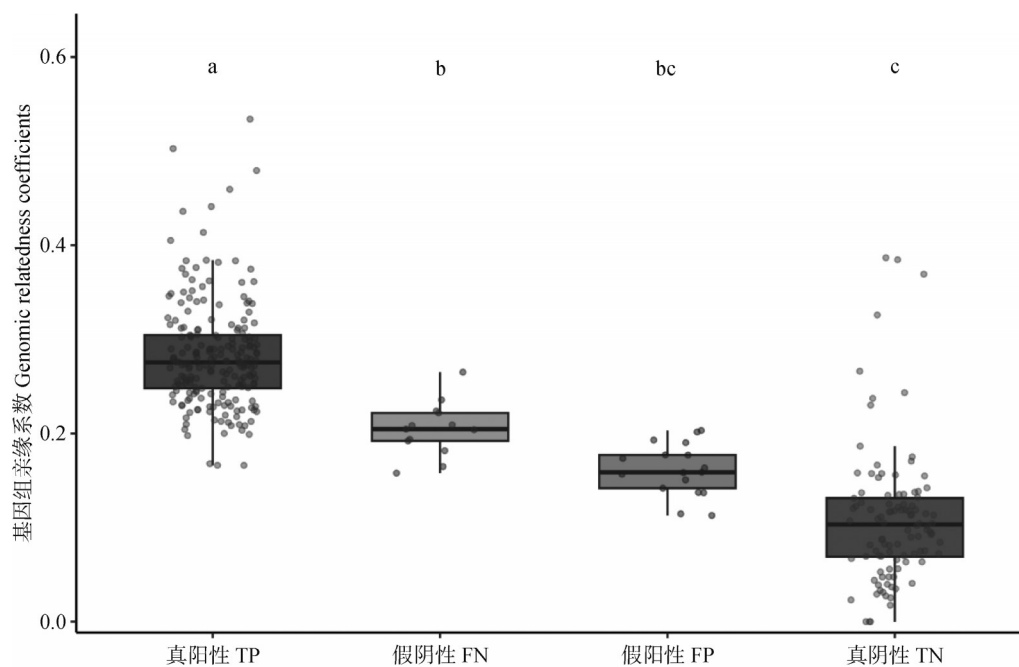


图3 不同识别类型下个体与候选外祖父的基因组亲缘系数分布图

Fig. 3 Distribution of genomic relatedness coefficients between individuals and candidate maternal grandsires under different identification categories

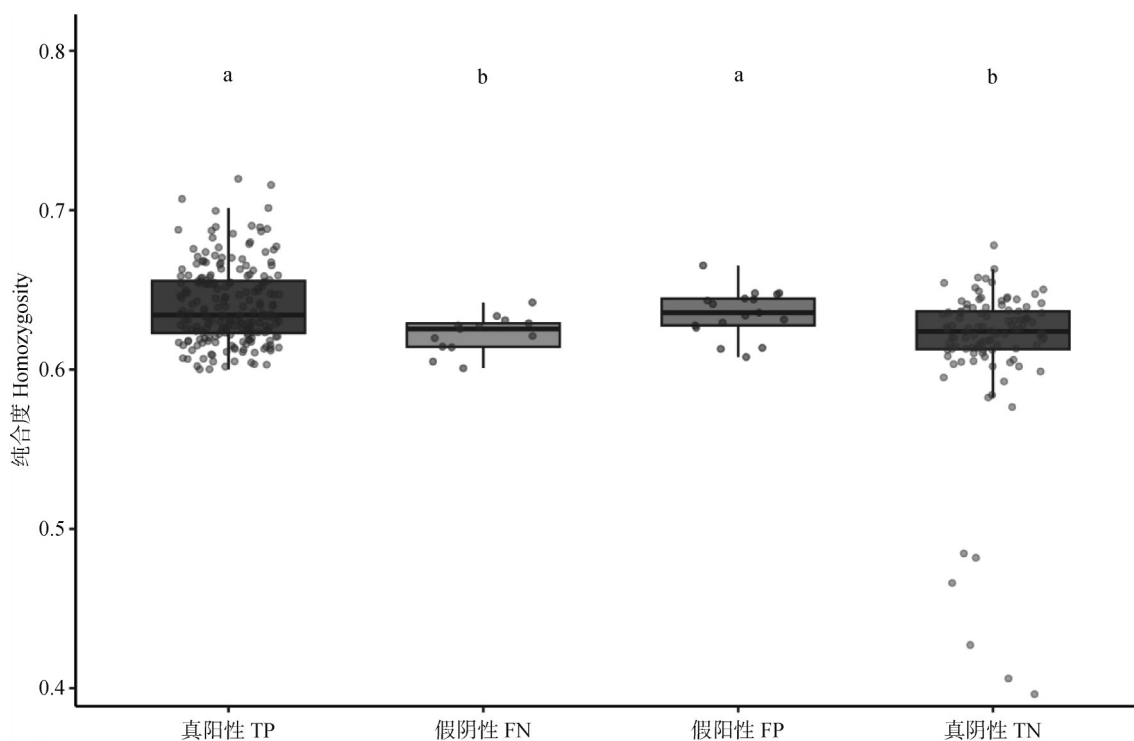


图4 不同识别类型下待测个体的纯合度分布图

Fig. 4 Distribution of homozygosity for the tested individuals under different identification categories

3 讨论

本研究基于奶牛的真实生产群体,采用分段单倍型匹配的方法开展外祖父鉴定,系统评估了芯片密度、单倍型窗口长度及群体遗传背景对系谱外祖父鉴定准确性的影响,旨在为单倍型鉴定方法的优化与实际应用提供理论依据和实践参考。

本研究中,该方法在大多数组合下的最高鉴定准确率均超过90%,整体表现出较高的准确性和稳健性,验证了其在大规模奶牛生产群体中的应用可行性与推广潜力。尤其在低密度芯片条件下,5K芯片 $\times$ 10个SNPs窗口长度组合的最高准确率达到93.9%,与VanRaden等^[19]在3K芯片条件下报告的93%结果相近,表明即便在分型信息有限的情况下,通过合理设置窗口参数,仍可实现较高的外祖父鉴定精度,满足实际生产中的应用需求。在中高密度芯片条件下,50K芯片 $\times$ 75个SNPs组合的准确率为96.2%,略低于VanRaden等^[19]报道的98%,但当窗口长度进一步优化为35个SNPs时,准确率提升至97.4%,显示该方法在分型信息丰富的条件下具备良好的鉴定能力与准确性。需要指出的是,本研究所设定的窗口长度范围仍有限,部分芯片密度条件下的理论最优组合可能尚未完全覆盖。此外,多数组合的最优匹配阈值与VanRaden等^[19]提出的35%经验值存在明显差异,提示需要根据自身群体的数据特征灵活调整阈值以进一步提升方法的适应性和准确性。

为深入探究芯片密度与单倍型窗口长度对外祖父(MGS)鉴定效果的影响,本研究分析发现不同窗口长度对芯片密度的响应存在差异,在短窗口(5、10个SNPs)条件下,4种芯片密度(5K、10K、20K、50K)的阳性与阴性组HMR分布均具备良好区分度,且整体准确性较高,因而进一步提升芯片密度所带来的边际效益较低。而中长窗口(35、75、120个SNPs)基础准确率偏低,分布分层较弱,对芯片密度提升更为敏感,表现出明显的准确率提升。表明在低成本密度情况下较短窗口的分段更具优势,而中长窗口对密度变化更为敏感,适用于中高密度芯片下的精细化判定。当芯片密度进一步增加至50K时,长窗口(75、120个SNPs)组合下的准确率仍有提升,而其他窗口组合下的准确率略有下降,但整体准确性仍维持在较高水平。分析认为,短窗口片段内各位点之间的平均遗传距离较近,在

高密度芯片条件下,连锁不平衡(LD)进一步增强,导致部分单倍型片段在群体中高度一致从而削弱了个体特异性信息的辨识力,使得无关个体在短区间内也可能出现随机匹配^[27-30]。相比之下,长窗口覆盖更长的遗传距离和更多的重组事件,导致片段一致性过低从而降低了个体间的可比性,同样难以实现有效匹配,但随着芯片密度的不断增加,长窗口内的连锁结构得到增强,片段的比对可比性和区分能力提升,最终优化了匹配效果^[31-35]。因此,为获得最佳鉴定效果,需在固定窗口长度下合理控制芯片密度,避免因密度过高或过低带来的准确性下降。总体来看,芯片密度越高,各窗口下的平均准确率越高,而窗口长度对准确性的影响呈现“先升后降”趋势,大多数组合在35个SNPs处达到最佳。此外,芯片密度与最佳判定阈值之间呈正相关关系,随着密度提高,匹配率上升,需同步提升判定阈值;而窗口长度越长,由于单倍型一致性降低,HMR整体水平下降,最优阈值相应偏低。

值得注意的是,HMR分布结果显示,随着芯片密度由5K提升至20K,部分组合的单倍型分辨能力和群体区分度明显增强,阳性验证组中部分个体呈现出独立的高HMR区域。在中长窗口条件下,这种分层现象愈发明显,分析认为,这与相关个体较高的近交水平有关,较长窗口更易覆盖近交所导致的连续纯合片段,从而维持较高的匹配率^[36]。相比之下,非近亲个体由于片段中包含更多重组事件,在长窗口条件下难以形成稳定的一致性匹配,导致其HMR分布趋近于阴性组。

本研究还分析了群体亲缘背景与近交水平对单倍型方法鉴定结果的潜在影响,为不同遗传背景下的应用提供参考。亲缘关系分析表明,亲缘关系与HMR呈现出高度的一致性,在一定程度上解释了该方法能够清晰反映个体间的亲缘关系,而观察TN组个体的亲缘系数虽整体较低,但仍存在极少部分亲缘系数较高的个体,且其对应的HMR仍较低,提示单倍型匹配率除亲缘关系外,可能还受到其他因素影响^[37-38]。进一步结合近交水平分析揭示了群体遗传背景对匹配效果的深层影响,进一步分析后发现TN组中纯合度极低的个体与此前分析的少数亲缘系数较高的个体高度重合,表明即使个体与候选外祖父存在较近的亲缘关系,若其基因组背景杂合程度较高,单倍型片段频繁重组也可能导致匹配效果减弱^[39-40]。因此,在实际应用中,应根据

群体亲缘结构与个体近交水平动态调整阈值,以提高方法在复杂群体背景下的鉴定效果^[41-42]。

需要指出的是,由于本研究中窗口长度设置有限,高密度芯片下的理论最优窗口可能尚未被充分覆盖,当前结果未必代表该条件下的最优准确性。此外,所采用的 Findhap 软件在进行单倍型推断时会融合群体聚类与系谱等先验信息进行迭代优化,因此试验群体的样本规模、基因组数据完整性及系谱准确性均可能对鉴定结果产生一定影响。鉴于上述多种因素影响,建议优先筛选具有亲子关系验证的祖孙关系作为参数优化训练集,可结合本研究所揭示的芯片密度、窗口长度及遗传背景对鉴定结果的影响趋势为参考,辅助匹配率阈值的合理设定,以进一步提升外祖父鉴定方法的准确性与适用性。

4 结 论

本研究以 6 117 头真实生产奶牛为对象,系统评估了分段单倍型匹配法在不同芯片密度、窗口长度及群体遗传背景下的外祖父鉴定性能。结果表明:1)该方法在绝大多数组合下均可实现超过 90% 的鉴定准确率,其中 20K×35 个 SNPs 组合表现最佳,准确率达 98%,展现出良好的准确性与稳健性;2)芯片密度与窗口长度对鉴定效果存在显著影响,芯片密度越高,平均准确率整体越高,而窗口长度对准确性的影响呈现“先升后降”趋势,最优匹配阈值随芯片密度升高而增大,随窗口长度延长而降低。通过合理搭配芯片密度与窗口长度,并设定合适的阈值,不仅可优化鉴定效果,还能在降低成本的同时实现高准确率;3)群体亲缘关系与近交水平显著影响 HMR 分布与鉴定效果,较高的亲缘关系和近交水平有助于提高匹配率,但也增加了非真实祖孙间被误判为阳性的风险;而亲缘关系和近交水平较低,则更易导致真实外祖父的漏检。因此,通过结合群体遗传背景合理搭配芯片密度与窗口长度,并设定适宜的阈值,不仅可优化鉴定效果,还能在降低成本的同时实现高准确率。该方法为完善奶牛母系谱系提供了有效途径,为奶牛遗传评估和育种决策提供可靠支持。

参考文献 (References):

[1] HU L R, LI D, CHU Q, et al. Selection and implementation of single nucleotide polymorphism

markers for parentage analysis in crossbred cattle population [J]. *Animal*, 2021, 15(1):100066.

[2] WIGGANS G R, CARRILLO J A. Genomic selection in United States dairy cattle [J]. *Front Genet*, 2022, 13:994466.

[3] 孙东晓,张胜利,张 勤,等. 我国奶牛基因组选择技术应用进展[J]. 畜牧兽医学报, 2023, 54(10):4028-4039.

SUN D X, ZHANG S L, ZHANG Q, et al. Application progress on genomic selection technology for dairy cattle in China [J]. *Acta Veterinaria et Zootechnica Sinica*, 2023, 54 (10) : 4028-4039. (in Chinese)

[4] GHAVI HOSSEIN-ZADEH N. An overview of recent technological developments in bovine genomics [J]. *Vet Anim Sci*, 2024, 25:100382.

[5] GARCÍA-RUIZ A, WIGGANS G R, RUIZ-LÓPEZ F J. Pedigree verification and parentage assignment using genomic information in the Mexican Holstein population[J]. *J Dairy Sci*, 2019, 102(2):1806-1810.

[6] HIMMELBAUER J, SCHWARZENBACHER H, FUERST C, et al. Exploring unknown parent groups and metafounders in single-step genomic best linear unbiased prediction: Insights from a simulated cattle population [J]. *J Dairy Sci*, 2024, 107 (10) : 8170-8192.

[7] PIMENTEL E C G, EDEL C, EMMERLING R, et al. How pedigree errors affect genetic evaluations and validation statistics[J]. *J Dairy Sci*, 2024, 107(6) : 3716-3723.

[8] WELLMANN R, ROLFES A, RENSING S, et al. Economic benefits of herd genotyping and using sexed semen for pure and beef-on-dairy breeding in dairy herds[J]. *J Dairy Sci*, 2024, 107(5):2983-2998.

[9] STRABEL T. Association of pedigree indexes and genomic breeding values with the performance of Polish Holstein-Friesian cows[J]. *J Appl Genet*, 2025, 66(1):207-218.

[10] WIGGANS G R, VANRADEN P M, BACHELLER L R. Methods for discovering and validating relationships among genotyped animals[C]. Auckland, New Zealand: Interbull Bulletin, 2018, 53:27-30.

[11] MASUDA Y, VANRADEN P M, TSURUTA S, et al. Invited review: unknown-parent groups and metafounders in single-step genomic BLUP [J]. *J Dairy Sci*, 2022, 105(2):923-939.

[12] BRADFORD H L, MASUDA Y, VANRADEN P

- M, et al. Modeling missing pedigree in single-step genomic BLUP [J]. *J Dairy Sci*, 2019, 102(3): 2336-2346.
- [13] LÓPEZ-CORREA R D, LEGARRA A, AGUILAR I. Modeling missing pedigree with metafounders and validating single-step genomic predictions in a small dairy cattle population with a great influence of foreign genetics [J]. *J Dairy Sci*, 2024, 107(7): 4685-4692.
- [14] NANI J P, BACHELLER L R, COLE J B, et al. Discovering ancestors and connecting relatives in large genomic databases [J]. *J Dairy Sci*, 2020, 103(2): 1729-1734.
- [15] STRUCKEN E M, LEE S H, LEE H K, et al. How many markers are enough? Factors influencing parentage testing in different livestock populations [J]. *J Anim Breed Genet*, 2016, 133(1): 13-23.
- [16] MCCLURE M C, MCCARTHY J, FLYNN P, et al. SNP data quality control in a national beef and dairy cattle system and highly accurate SNP-based parentage verification and identification [J]. *Front Genet*, 2018, 9: 84.
- [17] GEBREHIWOT N Z, STRUCKEN E M, MARSHALL K, et al. SNP panels for the estimation of dairy breed proportion and parentage assignment in African crossbred dairy cattle [J]. *Genet Sel Evol*, 2021, 53(1): 21.
- [18] KOMIYA R, OGAWA S, AONUMA T, et al. Performance of using opposing homozygotes for paternity testing in Japanese Black cattle [J]. *J Anim Breed Genet*, 2022, 139(1): 113-124.
- [19] VANRADEN P M, COOPER T A, WIGGANS G R, et al. Confirmation and discovery of maternal grandsires and great-grandsires in dairy cattle [J]. *J Dairy Sci*, 2013, 96(3): 1874-1879.
- [20] VANRADEN P M, O'CONNELL J R, WIGGANS G R, et al. Genomic evaluations with many more genotypes [J]. *Genet Sel Evol*, 2011, 43(1): 10.
- [21] VANRADEN P M, SUN C, O'CONNELL J R. Fast imputation using medium or low-coverage sequence data [J]. *BMC Genet*, 2015, 16: 82.
- [22] 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: 7.
- [23] COUTINHO-LIMA D, DREGER D L, DOADRIO I, et al. Multiple ancestries and shared gene flow among modern livestock guarding dogs [J]. *iScience*, 2024, 27(8): 110396.
- [24] LUŠTREK B, ŠIMON M, TURK K, et al. Comparing genomic and pedigree inbreeding coefficients in the Slovenian Lipizzan horse as a case study for small closed populations [J]. *Animals (Basel)*, 2025, 15(19): 2774.
- [25] LUIGI-SIERRA M G, FERNÁNDEZ A, MARTÍNEZ A, et al. Genomic patterns of homozygosity and inbreeding depression in Murciano-Granadina goats [J]. *J Anim Sci Biotechnol*, 2022, 13(1): 35.
- [26] HU M, JIANG H, LAI W, et al. Assessing genomic diversity and signatures of selection in Chinese Red Steppe cattle using high-density SNP array [J]. *Animals (Basel)*, 2023, 13(10): 1717.
- [27] KHATKAR M S, NICHOLAS F W, COLLINS A R, et al. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel [J]. *BMC Genomics*, 2008, 9: 187.
- [28] THOMPSON E A. Identity by descent: variation in meiosis, across genomes, and in populations [J]. *Genetics*, 2013, 194(2): 301-326.
- [29] STRUCKEN E M, AL-MAMUN H A, ESQUIVELZETA-RABELL C, et al. Genetic tests for estimating dairy breed proportion and parentage assignment in East African crossbred cattle [J]. *Genet Sel Evol*, 2017, 49(1): 67.
- [30] NAIT SAADA J, KALANTZIS G, SHYR D, et al. Identity-by-descent detection across 487, 409 British samples reveals fine-scale population structure and ultra-rare variant associations [J]. *Nat Commun*, 2020, 11(1): 6130.
- [31] BOHMANOVA J, SARGOLZAEI M, SCHENKEL F S. Characteristics of linkage disequilibrium in North American Holsteins [J]. *BMC Genomics*, 2010, 11: 421.
- [32] BROWNING B L, BROWNING S R. A fast, powerful method for detecting identity by descent [J]. *Am J Hum Genet*, 2011, 88(2): 173-182.
- [33] SU S Y, KASBERGER J, BARANZINI S, et al. Detection of identity by descent using next-generation whole genome sequencing data [J]. *BMC Bioinformatics*, 2012, 13: 121.
- [34] FERDOSI M H, HENSHALL J, TIER B. Study of the optimum haplotype length to build genomic relationship matrices [J]. *Genet Sel Evol*, 2016, 48(1): 75.
- [35] CEBALLOS F C, JOSHI P K, CLARK D W, et al. Runs of homozygosity: windows into population

- history and trait architecture [J]. *Nat Rev Genet*, 2018, 19(4):220-234.
- [36] IACOLINA L, STRONEN A V, PERTOLDI C, et al. Novel graphical analyses of runs of homozygosity among species and livestock breeds [J]. *Int J Genomics*, 2016, 2016:2152847.
- [37] EDWARDS D. Two molecular measures of relatedness based on haplotype sharing [J]. *BMC Bioinformatics*, 2015, 16:383.
- [38] GRASHEI K E, ØDEGÅRD J, MEUWISSEN T H E. Using genomic relationship likelihood for parentage assignment [J]. *Genet Sel Evol*, 2018, 50(1):26.
- [39] MARRAS G, GASPA G, SORBOLINI S, et al. Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy [J]. *Anim Genet*, 2015, 46(2):110-121.
- [40] FORUTAN M, ANSARI MAHYARI S, BAES C, et al. Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle [J]. *BMC Genomics*, 2018, 19(1):98.
- [41] PANETTO J C C, MACHADO M A, SILVA M V G B, et al. Parentage assignment using SNP markers, inbreeding and population size for the Brazilian Red Sindhi cattle [J]. *Livest Sci*, 2017, 204:33-38.
- [42] DODDS K G, MCEWAN J C, BRAUNING R, et al. Exclusion and genomic relatedness methods for assignment of parentage using genotyping-by-sequencing data [J]. *G3 (Bethesda)*, 2019, 9(10):3239-3247.
- (编辑 郭云雁)