

曾浩南, 钟展明, 徐志婷, 等. 3 款猪 50K SNP 芯片基因型填充至序列数据的效果评估[J]. 华南农业大学学报, 2022, 43(4): 10-15.
ZENG Haonan, ZHONG Zhanming, XU Zhiting, et al. Evaluation on genotype imputation performance of three porcine 50K SNP chips from chip data to sequencing data[J]. Journal of South China Agricultural University, 2022, 43(4): 10-15.

3 款猪 50K SNP 芯片基因型填充至 序列数据的效果评估

曾浩南, 钟展明, 徐志婷, 滕金言, 袁晓龙, 李加琪, 张 哲
(华南农业大学 动物科学学院/广东省农业动物基因组学与分子育种重点实验室/
国家生猪种业工程技术研究中心, 广东 广州 510642)

摘要:【目的】利用猪 50K SNP(Single nucleotide polymorphisms) 芯片开展基因组育种已经得到了广泛的应用与认可。基因型填充可在不增加基因型检测成本的前提下大幅提高基因型数据量, 有利于开展复杂性状的遗传解析与遗传评估。本研究旨在评估 3 款猪 SNP 芯片基因型填充至序列数据的填充效果。【方法】选用 3 款芯片共同检测的 48 头杜洛克猪群体作为填充的目标群体, 260 头猪的全基因组测序数据作为参考群体, 使用 Beagle5.1 软件进行基因型填充, 对比 3 款不同猪 SNP 芯片纽勤 50K、中芯一号 50K 和液相 50K 基因型填充至序列数据的填充效果。【结果】3 款芯片原始的 SNP 数分别为 50 697、57 466 和 50 885 个。填充至序列后, 未质控时位点填充准确性(基因型一致性) 分别为 0.886、0.886 和 0.898, 质控过滤 DR^2 (Dosage R-squared) <0.95 的位点后, 填充准确性(基因型一致性) 分别提升至 0.974、0.976 和 0.969, 位点数分别为 3393066、3 139095 和 3320627 个。【结论】不同芯片基因型填充至序列数据具有可行性, 通过基因型填充可获得高质量的高密度基因型数据, 可为后续的育种应用研究打下基础。

关键词: 猪; SNP 芯片; 基因型填充; 序列数据; 基因型一致性
中图分类号: S828.2 文献标志码: A 文章编号: 1001-411X(2022)04-0010-06

Evaluation on genotype imputation performance of three porcine 50K SNP chips from chip data to sequencing data

ZENG Haonan, ZHONG Zhanming, XU Zhiting, TENG Jinyan, YUAN Xiaolong, LI Jiaqi, ZHANG Zhe
(College of Animal Science, South China Agricultural University/Guangdong Provincial Key Laboratory of
Agro-animal Genomics and Molecular Breeding/National Engineering Research
Center for Breeding Swine Industry, Guangzhou 510642, China)

Abstract: 【Objective】Porcine 50K SNP (single nucleotide polymorphisms) chips have been widely used in pig genomic breeding. Meanwhile, genotype imputation can significantly increase the amount of genotype data without increasing the cost of sequencing, which facilitates genetic resolution and genetic evaluation of complex traits. This study was aimed to evaluate the genotype imputation performance from genotype to sequence data of three porcine SNP chips. 【Method】A total of 48 Duroc pigs with three kinds of porcine SNP chips were used as target panel to evaluate the genotype imputation accuracy. A total of 260 pigs with whole genome sequencing

data formed a reference panel for genotype imputation. The genotype imputation was performed using Beagle5.1 software to compare the imputation effect of Geneseek 50K, Zhongxin I 50K and Liquid 50K. 【Result】 The numbers of original SNPs in three kinds of chips were 50697, 57466 and 50885 respectively. The imputation accuracies (genotype consistencies) were 0.886, 0.886 and 0.898 respectively after imputation without any quality control. After filtering the imputed SNPs with low reliability DR^2 (Dosage R-squared) <0.95 , the imputation accuracies (genotype consistencies) of three kinds of chips were up to 0.974, 0.976 and 0.969 respectively, and the numbers of remaining SNPs were 3393066, 3139095 and 3320627 respectively. 【Conclusion】 Genotype data from the three types of porcine SNP chips can be imputed to sequence data with a high imputation accuracy. This study provides useful reference for subsequent breeding application research.

Key words: Pig; SNP chip; Genotype imputation; Sequence data; Genotype consistency

单核苷酸多态性 (Single nucleotide polymorphisms, SNP) 是指由于单个核苷酸发生变异引起的基因组水平上 DNA 序列的多态性, 具有遗传稳定性强、位点数量多且分布广, 易于大规模自动化检测的优势, SNP 芯片作为检测 SNP 的重要手段, 已被广泛应用于生物、医学、农学等领域^[1]。在猪育种中, 相对低密度芯片的位点信息不足和高密度芯片的价格高昂, 中密度 50K 芯片成为了猪基因组选择的主流芯片类型。其中, Neogen 子公司 Geneseek 研发的纽勤 50K(Porcine SNP 50K beadchip)、江西农业大学研发的中芯一号 50K(KPS Porcine breeding chip v2) 及中国农业大学研发的液相 50K(GenoBaits porcine SNP 50K) 为国内常见的 3 款猪 50K SNP 芯片。3 款芯片设计的原理不同, 纽勤 50K 和中芯一号 50K 均是基于 Illumina 平台研发的固相芯片, 液相 50K 芯片是基于靶向测序基因型分型技术^[2] 研发的芯片。

全基因组测序数据 (Whole genome sequencing, WGS) 相比芯片数据记录了更完整的遗传信息, 但由于直接测序成本较高, 因此人们常用基因型填充技术将低密度的芯片数据填充至序列数据。基因型填充是一项根据参考群体的基因型信息推断目标群体缺失基因型的技术, 其能够充分利用现有的 SNP 芯片信息将低密度数据准确填充至高密度, 实现了在节约育种成本的基础上提高育种准确性^[3-4]。现有的基因型填充软件有很多, 常用的有 Beagle^[5-6]、Impute2^[7]、Findhap^[8]、AlphaImpute^[9]、FAMHAP^[10]、FImpute^[11] 等。基因型填充技术现已成为一种常规、有效的工具, 被广泛应用于动物遗传育种领域, 在猪的相关研究中, 基于填充后的数据有助于筛选新的候选基因^[12]、提升基因组选择的准确性, 并降低生产的成本^[13-15]。同样, 在牛^[16-17]、羊^[18]、鸡^[19] 的遗传育种研究中基因型填充也起到了重要的作用。

一方面, 基因型填充的效果受到许多数据如参考群体大小、参考群体与目标群体之间的遗传关系、最小等位基因频率 (Minor allele frequency, MAF) 等的影响; 另一方面, 不同填充软件计算的质控指标对填充准确性也有一定的影响, 如 Beagle^[5-6] 软件的 DR^2 (Dosage R-squared)、Impute^[7] 软件的 INFO 等。

此前, 国内研究团队对 3 款芯片之间相互填充的效果进行了评估^[20], 但对 3 款芯片填充至序列数据的效果鲜有报道, 因此本研究旨在评估 3 款猪 50K SNP 芯片基因型填充至序列数据的效果, 以期为我国猪遗传研究和分子育种工作提供参考和借鉴。

1 材料与方法

1.1 芯片群体

本研究所用的 48 头杜洛克猪的芯片群体来自广西某育种场, 该群体在纽勤 50K、中芯一号 50K 及液相 50K 芯片的位点数分别为 50 697、57466 和 50885, 具体的位点重叠情况见图 1a。另外, 在芯片重叠的位点里, 本研究在等位基因精确匹配后评估了 3 款芯片两两之间的一致性与相关性。

芯片数据利用根据以下步骤进行质量控制: 1) 剔除位置重复或未知的位点; 2) 剔除非常染色体位点; 3) 利用 Conform-gt 软件校正正反链并去除参考群以外的位点。质控后纽勤 50K、中芯一号 50K 及液相 50K 芯片剩余位点数分别为 31 756、29 469 和 42 311, 分别占原始芯片数据的 62.6%、51.3% 和 83.2%, 质控后具体的位点重叠情况见图 1b。

1.2 参考群体

本研究使用的全基因组测序参考群体共

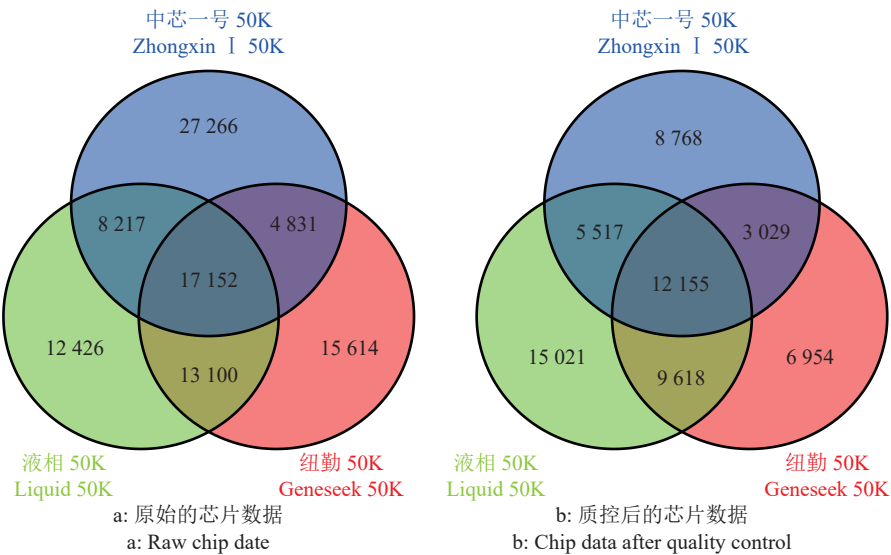


图 1 3 款芯片之间的位点分布
Fig. 1 Distribution of loci among three chips

260 头猪, 其中 50 头来自福建某种猪场的杜洛克, 210 头来自华南地区的 21 个地方猪品种。该群体平均测序深度为 14.53×, 平均测序覆盖度为 98.85%, 经过 MAF<0.01 的质量控制后, 本研究使用 Beagle5.1 对其进行单倍型分型 (Phasing), 最后参考群体的 SNP 位点数为 31 407 555 个。

1.3 基因型填充方法及准确性评估

本研究使用 Beagle5.1 软件对芯片数据的 18 条常染色体进行基因型填充。评估填充准确性的步骤如下: 1) 在填充前将芯片数据有序不重复地缺失 5%, 共缺失 20 次; 2) 根据保留的 95% 芯片位点填充缺失的 5% 芯片位点, 共填充 20 次; 3) 将 20 次填充后的结果合并在一起, 从位点水平上以缺失前与填充后的基因型一致性与相关性作为基因型填充准确性的评估指标。其中, 基因型一致性指的是基因型完全一致的个数占总基因型个数的比例, 而基因型相关性指的是将基因型转换为 0、1、2 编码方式后计算基因型之间的皮尔逊相关系数。

相比基因型一致性, 基因型相关性更能反映准确性整体的变化规律, 因此本研究在研究 MAF 与 DR² 对填充准确性的影响时, 均以基因型相关性作为填充准确性的指标。

2 结果与分析

2.1 芯片之间重叠位点的基因型一致性与相关性

表 1 展示了 3 款芯片之间位置与等位基因信息完全重叠的位点基因型一致性与相关性情况。从表 1 中可以看出, 3 款芯片重叠的位点基本保持一致, 其中 2 款固相芯片之间的一致性最高, 达到了

表 1 芯片之间重叠位点的基因型一致性与相关性
Table 1 The consistency and correlation of overlapping loci among three chips

芯片-芯片 Chip-chip	一致性 Consistency	相关性 Correlation
纽勤50K-中芯一号50K Geneseek 50K-Zhongxin I 50K	0.999	0.996
纽勤50K-液相50K Geneseek 50K - Liquid 50K	0.991	0.985
中芯一号50K-液相50K Zhongxin I 50K-Liquid 50K	0.991	0.987

0.999; 2 款固相芯片与液相 50K 芯片之间的一致性略低, 为 0.991。

2.2 芯片填充至序列数据的填充准确性

表 2 展示了 3 款芯片基因型填充后未进行任何质控处理的填充准确性情况。从表 2 中可以看出液相 50K 芯片的位点基因型一致性最高, 达到 0.898, 纽勤 50K 的位点基因型相关性最高, 达到 0.828, 此时, 3 款芯片填充后的位点数均为 31 407 555。

2.3 MAF 与 DR² 对填充准确性的影响

本研究将 3 款芯片合并为一个整体, 研究 MAF 对填充准确性 (位点的基因型相关性) 的影响, 结果见图 2a。从整体可以看出, 随着 MAF 的增加, 填充的准确性显示出上升的趋势; 当 MAF<0.1 时, 随着 MAF 的增加, 填充准确性得到了显著的提升, 而当 MAF>0.1 时, 填充的准确性也有提升的趋势, 但提升的程度趋于平缓。其中, 质控过滤 MAF<0.1 的位点时, 纽勤 50K、中芯一号 50K 及液

表 2 3 款芯片基因型填充至序列数据的填充准确性¹⁾

芯片 Chip	未质控 No quality control	质控标准 Quality control standard		
		MAF ≥ 0.1	DR ² ≥ 0.8	DR ² ≥ 0.95
纽勤50K Geneseek 50K	0.886(0.828)	0.873(0.838)	0.938(0.917)	0.974(0.966)
中芯一号50K Zhongxin I 50K	0.886(0.823)	0.876(0.835)	0.944(0.918)	0.976(0.959)
液相50K Liquid 50K	0.898(0.814)	0.866(0.825)	0.930(0.909)	0.969(0.960)

1)表中数据为位点基因型一致性(基因型相关性);基因型一致性指的是基因型完全一致的个数占总基因型个数的比例;基因型相关性用将基因型转换为0、1、2剂量编码后基因型之间的皮尔逊相关系数来表示

1) Data in the table are genotype consistencies (genotype correlations) of loci; Genotype consistency refers to the proportion of the number of completely consistent genotypes in the total number of genotypes; Genotype correlation is represented by the Pearson correlation coefficient between the genotypes after converting genotype to dosage encoding of 0, 1, and 2

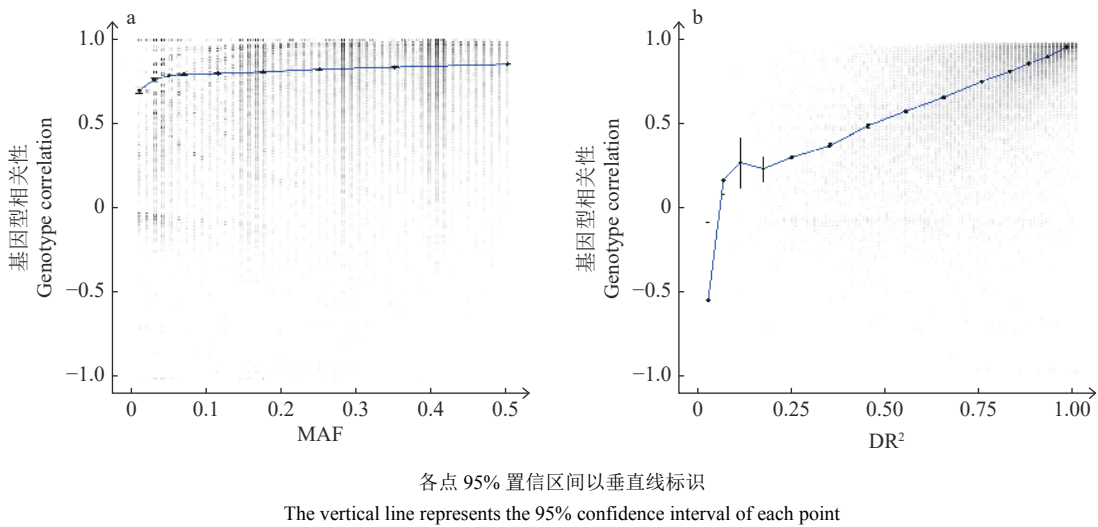


图 2 MAF、DR² 与填充准确性 (基因型相关性) 的分布

Fig. 2 Distribution of MAF, DR² and imputation accuracy (genotype correlation)

相 50K 这 3 款芯片的准确性分别为 0.838、0.835 和 0.825, 而此时剩余的位点数分别为 79 556 801、8202 065 和 7717 077。

DR² 可以作为 Beagle^[5-6] 软件基因型填充后的质控指标, 3 款芯片整体的 DR² 对填充准确性的影响结果见图 2b。从图 2b 中可以明显看到 DR² 与填充准确性存在较强的正相关关系, 其中, 质控过滤 DR²<0.4 的位点后填充的准确性增长趋势较为缓慢, 当逐步剔除 0.4≤DR²≤1 的位点时, 3 款芯片的填充准确性得到了较有效的提升。相比 DR²≥0.4 的情况, DR²≥0.6 时, 纽勤 50K、中芯一号 50K 和液相 50K 芯片的填充准确性分别提升了 3.84%、4.16% 和 3.92%; DR²≥0.8 时, 3 款芯片的填充准确性分别提升了 9.17%、9.44% 和 9.88%; 当 DR²≥0.95 时, 3 款芯片的填充准确性分别提升了 14.95%、14.37% 和 16.11%, 此时的准确性分别为 0.966、0.959 和 0.960, 剩余的位点数为 3 393 066、3 139 095 和 3 320 627。

3 讨论与结论

本研究展示了国内 3 款常用的猪 50K SNP 芯片的基本情况与基因型填充至序列数据的结果。从芯片的基本情况来看, 3 款芯片均有较高的 SNP 检出率及重叠位点的一致率, 3 款芯片的 SNP 位点数均达到 50K 芯片的标准。在本研究中, 中芯一号 50K 芯片虽然原始位点数较多, 但非多态的位点也较多, 导致填充前质控位点后剩余的位点数较少, 一方面的原因是本研究使用的芯片数据个体数偏少导致位点整体的多态率较低, 另一方面可能是中芯一号 50K 芯片位置的设计原则与其余 2 款芯片的有较大的出入, 但从结果上看这并没有导致中芯一号 50K 芯片基因型填充的效果弱于其余 2 款芯片。

从芯片的基因型填充的结果上看, 3 款芯片的原始填充效果并没有明显的差异。此外, 在本研究中 3 款芯片填充后的位点基因型一致性均值为

0.890, 尚未达到很高的水平。一方面, 已有研究证实参考群规模的大小会影响填充的效果^[21-22], 本研究使用的参考群体个体数为 260, 在参考群体规模上一定程度地限制了填充的效果; 另一方面, 本研究参考群体里杜洛克个体数仅为 50, 个体数少导致参考群体的杜洛克单倍型信息并不能很好地囊括整个品种, 进而填充的效果降低。

MAF 也是影响填充准确性的一个重要因素, 据报道, MAF<0.05 时, MAF 每提升 0.01 填充准确性都会得到较大的提升, 而 MAF≈0.1 已经与最高点没有明显的差异^[23-27], 在本研究中, MAF 的趋势也与前人的结果基本符合。

DR² 作为 Beagle^[5-6] 填充的质控指标, 根据 DR² 进行基因型填充后, 位点的质控可以最直接地提升基因型填充的准确性, 从而提升全基因组关联分析及基因组选择等下游分析的效果^[28-31]。从本研究的结果看, 当 DR²≥0.95 时, 纽勤 50K、中芯一号 50K 及液相 50K 芯片的填充效果很好 (基因型一致性与相关性均高于 0.950), 但是位点数也由原来的约 4 000 万减少到约 300 万。在基因组选择领域, 一方面位点数的减少可能会导致位点变异解释的遗传力降低, 从而影响基因组选择的效果^[32], 另一方面位点错误率升高也会降低基因组选择的效果^[33], 因此在实际应用中, DR² 的阈值选取需要研究者根据自身需求进行考量。需要注意的是, DR² 和填充准确性的增长趋势与 Beagle 设置的参数有效群体大小 (ne) 有关, 默认的参数较大 (1 000 000), 而畜禽育种的有效群体大小相比人类普遍较低, 因此基因型填充时需要根据参考群体的规模适当进行该参数的调整, 以适应更多畜禽育种的群体情况, 这有利于提升填充效果^[5-6]。

综上所述, 本研究评估了 3 款猪 SNP 芯片基因型填充至序列数据的效果, 结果表明 3 款芯片基因型填充至序列数据的策略可行, 均可获得较高的填充准确性。

参考文献:

[1] 唐立群, 肖层林, 王伟平. SNP 分子标记的研究及其应用进展[J]. *中国农学通报*, 2012, 28(12): 154-158.

[2] 徐云碧, 杨泉女, 郑洪建, 等. 靶向测序基因型检测 (GBTS) 技术及其应用[J]. *中国农业科学*, 2020, 53(15): 2983-3004.

[3] 何桑, 丁向东, 张勤. 基因型填充方法介绍及比较[J]. *中国畜牧杂志*, 2013, 49(23): 95-100.

[4] 叶绍潘. 基于全基因组测序数据的基因型填充准确性研究[D]. 广州: 华南农业大学, 2017.

[5] BROWNING B L, ZHOU Y, BROWNING S R. A one-

penny imputed genome from next-generation reference panels[J]. *American Journal of Human Genetics*, 2018, 103(3): 338-348.

[6] BROWNING S R, BROWNING B L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering[J]. *American Journal of Human Genetics*, 2007, 81(5): 1084-1097.

[7] HOWIE B N, DONNELLY P, MARCHINI J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies[J]. *PLoS Genetics*, 2009, 5(6): e1000529.

[8] VANRADEN P M, SUN C, O'CONNELL J R. Fast imputation using medium or low-coverage sequence data[J]. *BMC Genetics*, 2015, 16: 82.

[9] HICKEY J M, KINGHORN B P, TIER B, et al. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation[J]. *Genetics Selection Evolution*, 2012, 44(1): 9.

[10] BECKER T, KNAPP M. Maximum-likelihood estimation of haplotype frequencies in nuclear families[J]. *Genetic Epidemiology*, 2004, 27(1): 21-32.

[11] SARGOLZAEI M, CHESNAIS J P, SCHENKEL F S. A new approach for efficient genotype imputation using information from relatives[J]. *BMC Genomics*, 2014, 15: 478.

[12] 汪楷庭, 付璐, 孟庆利, 等. 基于填充测序数据的大白猪繁殖性状全基因组关联分析[C]//中国畜牧兽医学会. 第三届中国猪业科技大会暨中国畜牧兽医学会 2019 年学术年会论文集. 青岛: 中国畜牧兽医学会, 2019: 55.

[13] CLEVELAND M A, HICKEY J M. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation[J]. *Journal of Animal Science*, 2013, 91(8): 3583-3592.

[14] ZHANG C, KEMP R A, STOTHARD P, et al. Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants[J]. *Genetics Selection Evolution*, 2018, 50(1): 14.

[15] GROSSI D A, BRITO L F, JAFARIKIA M, et al. Genotype imputation from various low-density SNP panels and its impact on accuracy of genomic breeding values in pigs[J]. *Animal: An International Journal of Animal Bioscience*, 2018, 12(11): 2235-2245.

[16] ALILOO H, MRODE R, OKEYO A M, et al. The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa[J]. *Journal of Dairy Science*, 2018, 101(10): 9108-9127.

[17] IBEAGHA-AWEMU E M, PETERS S O, AKWANJI K A, et al. High density genome wide genotyping-by-sequencing and association identifies common and low frequency SNPs, and novel candidate genes influencing cow milk traits[J]. *Scientific Reports*, 2016, 6: 31109.

[18] TALOUARN E, BARDOU P, PALHIÈRE I, et al. Genome wide association analysis on semen volume and milk yield using different strategies of imputation to whole genome sequence in French dairy goats[J]. *BMC Genetics*, 2020, 21(1): 19.

[19] HUANG S, HE Y, YE S, et al. Genome-wide association study on chicken carcass traits using sequence data imputed from SNP array[J]. *Journal of Applied Genetics*, 2018, 59(3): 335-344.

[20] 邱奥, 王雪, 孟庆利, 等. 3 款猪 50K SNP 芯片基因型填充效果研究[J]. *中国畜牧杂志*, 2021, 57(S1): 33-38.

[21] BUTTY A M, SARGOLZAEI M, MIGLIOR F, et al. Optimizing selection of the reference population for genotype imputation from array to sequence variants[J]. *Frontiers in Genetics*, 2019, 10: 510.

[22] PAUSCH H, AIGNER B, EMMERLING R, et al. Imputation of high-density genotypes in the Fleckvieh cattle population[J]. *Genetics Selection Evolution*, 2013, 45(1): 3.

[23] DAETWYLER H D, CAPITAN A, PAUSCH H, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle[J]. *Nature Genetics*, 2014, 46(8): 858-865.

[24] MCCARTHY S, DAS S, KRETZSCHMAR W, et al. A reference panel of 64, 976 haplotypes for genotype imputation[J]. *Nature Genetics*, 2016, 48(10): 1279-1283.

[25] DAVIES R W, FLINT J, MYERS S, et al. Rapid genotype imputation from sequence without reference panels[J]. *Nature Genetics*, 2016, 48(8): 965-969.

[26] DAVIES R W, KUCKA M, SU D, et al. Rapid genotype imputation from sequence with reference panels[J]. *Nature Genetics*, 2021, 53(7): 1104-1111.

[27] RUBINACCI S, RIBEIRO D M, HOFMEISTER R J, et al. Efficient phasing and imputation of low-coverage sequencing data using large reference panels[J]. *Nature Genetics*, 2021, 53(1): 120-126.

[28] BOLORMAA S, GORE K, VAN DER WERF J H J, et al. Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy[J]. *Animal Genetics*, 2015, 46(5): 544-556.

[29] VAN DEN BERG I, BOICHARD D, LUND M S. Comparing power and precision of within-breed and multi-breed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds[J]. *Journal of Dairy Science*, 2016, 99(11): 8932-8945.

[30] PICCOLI M L, BRITO L F, BRACCINI J, et al. Genomic predictions for economically important traits in Brazilian Braford and Hereford beef cattle using true and imputed genotypes[J]. *BMC Genetics*, 2017, 18(1): 2.

[31] 王珏, 刘成琨, 刘德武, 等. 基于不同密度 SNP 芯片在杜洛克公猪中的全基因组选择效果分析[J]. *中国畜牧杂志*, 2019, 55(12): 75-79.

[32] DUFFLOCQ P, PÉREZ-ENCISO M, LHORENTE J P, et al. Accuracy of genomic predictions using different imputation error rates in aquaculture breeding programs: A simulation study[J]. *Aquaculture*, 2019, 503: 225-230.

[33] AKBARPOUR T, HOSSEIN-ZADEH N G, SHADPARVAR A A. Marker genotyping error effects on genomic predictions under different genetic architectures[J]. *Molecular Genetics and Genomics*, 2021, 296(1): 79-89.

【责任编辑 庄 延】