畜禽群体中基于SNP标记的亲子鉴定及亲本推断方法

罗元宇 吴 鹏 贺金龙 陈赞谋 张 豪 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与分子育种重点实验室/国家生猪种业工程中心,广州 510642

摘要 利用双亲高密度SNP标记信息,在畜禽群体中进行亲子鉴定及亲本推断新方法的计算机程序开发,并使用模拟的600 000个SNP标记对该程序进行测试。结果表明:对系谱正确性进行鉴定时,SNP标记数大于100即可达到100%的亲子鉴定准确率;对错误的系谱关系进行潜在亲子关系推断时,SNP标记数大于300可保证100%的推断准确率。标记平均MAF较低会降低亲子推断准确率。在程序运行效率方面,当使用50 000的SNP标记对1 000条错误率为10%的系谱进行亲子鉴定和推断时,共耗时332.87 s,且计算耗时与标记数及个体数呈线性变化。本研究基于孟德尔遗传原理,设计并开发了基于双亲及后代基因型的亲子鉴定及亲本推断程序,该方法运行速度快,操作简单,准确性高,值得在畜禽群体基因组相关研究方面进行推广应用。

关键词 亲子鉴定;亲本推断;SNP标记;孟德尔错误;准确率

中图分类号 S 813 **文献标识码** A **文章编号** 1000-2421(2016)05-0068-07

系谱信息在动植物遗传育种研究中起着重要作用,而实际生产中获得的系谱普遍存在错误。据报道,世界范围内奶牛的系谱错误率平均约为11%[1],在我国,系谱错误率则相对较高,天津为12%[2],北京则高达17%～21%[3]。除奶牛外,在兔[4]、大马哈鱼[5]、狗[6]和鹳[7]等物种都有系谱错误的研究报道。这些错误系谱的产生可能是因为系谱人工记录错误,或者因多次配种、混合输精或体外受精、胚胎移植等现代繁殖技术的使用导致系谱无法准确记录。在畜禽育种工作中,错误的系谱不仅会影响种畜亲子关系的确定,而且会减慢群体的遗传进展[8],从而对动物育种经济效益等造成重大损失。同时,错误系谱对畜禽群体遗传及育种研究也会产生负面影响,所以亲子鉴定方法在动物遗传育种生产及研究中具有十分重要的价值。

可用于亲子鉴定的标记物随着相关检测技术的发展而不断变化。血型和血液蛋白型[9-10]最早被用于亲子鉴定,但此技术因检测需采集大量血样,且当公牛去世后无法应用导致在实际生产中受到很大限制[11]。随着DNA检测技术的发展,小卫星[12-13]、微卫星(microsatellite)[14-15]以及单核苷酸多态(single nucleotide polymorphism,SNP)[16-17]等标记先后被用于亲子鉴定研究。其中,微卫星标记是目前亲子鉴定研究的主流标记,已成功用于马[18]、牛[19-21]等家畜的亲子鉴定。而SNP标记以突变率低、全基因组覆盖率高、遗传稳定性高、分型准确率高和检测成本低[22]等优势,受到了研究者们的青睐,已被广泛应用于人类及动植物遗传研究中。目前,用于亲子鉴定的软件和方法主要有KINSHIP[23]、PAPA[24]、FAMOZ[25]、Cervus[26]和EasyPC[17]等。这些软件及方法除了EasyPC能利用高密度SNP标记外,其他的都是利用少量SNP或微卫星标记。

近年来,除单一进行亲子鉴定的研究外,其他一些利用高密度SNP标记的研究,在开展研究之前都需要进行系谱校正以提高其结果的准确率,例如,全基因组关联分析[27-28]和基因组选择[29]等。在这些研究中,若使用微卫星标记进行系谱检验并不现实,而且上述软件多数采用相对复杂的算法来保证亲子鉴定结果的准确率,从而导致运行效率大大降低,当标记数目过多时可能无法运行[17]。本研究在前期

收稿日期:2015-10-30

基金项目:国家现代农业(生猪)产业技术体系项目(CARS-36);科技部科技基础性工作专项(2014FY120800);国家自然科学基金项目(31200925);广东省自然科学基金项目(2014A030313453)

罗元宇,硕士研究生. 研究方向:分子数量遗传学与动物育种. E-mail:yuanyuluo@163.com

通信作者:张 哲,博士,副教授. 研究方向:分子数量遗传学与动物育种. E-mail:zhezhang@scau.edu.cn

构建的基于单亲信息进行亲子推断方法的基础上，构建一种直接利用双亲与后代的全基因组 SNP 标记进行亲子鉴定及亲本推断的新方法，并用模拟数据对该方法的应用效果及影响因素进行详细讨论，旨在开发一种简易、高效、实用的利用双亲及后代基因型数据进行亲子鉴定及亲本推断的方法和程序。

1　材料与方法

1.1　群体数据模拟

本研究采用的数据均通过群体遗传数据模拟软件 GPOPSIM[30] 模拟生成。共模拟了 10 个世代，每个世代包含 1 000 个个体（即 1 000 条系谱），每世代均为半同胞-全同胞混合家系。本研究以其中的第 3、4 两个世代作为研究对象，共 2 000 条系谱。在程序运行效率和亲子鉴定、亲本推断结果的准确率等后续研究中，从整个系谱中随机选取部分系谱用于分析。

对基因型数据，共模拟了 30 条染色体，每条染色体包含 20 000 个标记，共 600 000 个标记，标记间呈均匀分布。在进行系谱分析之前，对模拟数据进行质量控制，最小等位基因频率（minor allele frequency，MAF）＜0.01 的 SNP 位点将被剔除。另外，数据模拟假定不存在无效等位基因，无基因型错误。

1.2　基于双亲信息的孟德尔错误率计算方法

基于孟德尔遗传定律，单个遗传位点的 2 个等位基因均以孟德尔遗传方式来自 2 个亲本。基于此规律，本方法将待检测个体及其疑似双亲进行配对，构成待检测亲子对，然后对每个待测亲子对的每一个双等位基因的遗传位点进行孟德尔遗传判定，具体判定规则见表 1。若待检验的位点共有 N 个，在疑似亲子对间，不符合孟德尔遗传定律的双亲与后代组合数（见表 1）有 N_{me} 个，则孟德尔错误率 $R_e = N_{me}/N$。另外，上述计算方法基于如下假定：(1) DNA 样品采集和基因型检测过程中无个体号记录错误；(2) 无基因型分型错误，或基因型分型错误率极低，但完全由随机因素造成。

基于上述孟德尔错误率计算规则，在群体中将随机个体对（无关或有关个体对）进行孟德尔错误率计算，获得群体中孟德尔错误率的经验分布后，对错误率划定阈值，即可判定对待检测系谱的正确性，即亲子鉴定；同时也可以此标准排除后代个体的疑似亲本，从而推断出其真实亲本，即亲本推断。基于前期结果，本研究选取的孟德尔错误率阈值为0.01。

表 1　基于双亲基因型的孟德尔错误判定规则

Table 1　Rules for Mendelian error determination with knownparents genotype

后代基因型 Genotype of offspring	双亲基因型组合 Combination of parents genotype								
	AA			Aa			aa		
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
AA	√	√	×	√	√	×	×	×	×
Aa	×	√	√	√	√	√	√	√	×
aa	×	×	×	×	√	√	×	√	√

注：“√”表示符合孟德尔遗传定律；“×”表示不符合孟德尔遗传定律。Note：The “√” shows that it is conformed to the Mendel’s law and the “×” shows that it is not conformed to the Mendel’s law.

1.3　程序设计及开发

利用本文“1.2”所述原理，使用 R 语言（http：www.r-project.org）进行程序开发，程序命名为 EasyPC（easy pedigree checking）。程序结构如图 1 所示。该程序所需输入文件为群体基因型文件和待鉴定系谱文件，输出亲子鉴定及亲本推断结果，包括鉴定正确系谱、鉴定错误系谱、推断正确系谱以及错误率图形等文件。程序设置了多个参数选项，可根据具体的需求来选择合适的运行参数。该程序的 R 代码及相应测试数据已免费在线共享至 https：//github.com/SCAU-AnimalGenetics/EasyPC。

1.4　程序验证

用本文“1.1”所述的模拟数据对上述方法进行检验。计算环境为：CPU 主频 3.2 GHz，内存 20.0 GB，Windows 7 操作系统。

错误系谱是将模拟的正确系谱在个体间随机进行调换生成，为研究系谱错误率的影响，生成多个梯度的系谱错误率，分别为 5%、10%、15%、20%、25%和 30%。其中，系谱错误率指错误的系谱数占总系谱数的比例。为研究标记数的影响，在 600 000 个 SNP 标记中，均匀挑选不同个数的 SNP，生成多个梯度SNP标记数据集，分别为100、200、300、500、

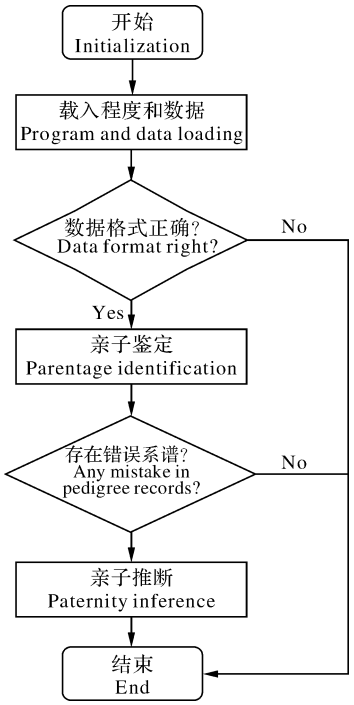

图1 EasyPC程序流程图

Fig.1 Structure chart of EasyPC program

1 000、5 000、10 000、15 000、20 000、25 000和50 000。为研究MAF的影响,从50 000数据集中,以0.2为阈值,按照MAF大小将标记分2组。为研究系谱数的影响,在1 000对系谱中,从100开始,按100递增至1 000,随机挑选相应数量的系谱。在上述不同条件下,研究该方法的准确率以及运行效率等。其中,除研究系谱数的影响外,其他分析均是从总系谱数中随机挑选1 000对系谱进行计算,且这些系谱中所包含的个体数平均为1 540(标准差为7)。

依据该方法的设计原理,为便于结果的记录和分析,我们将准确率分为亲子鉴定准确率和亲本推断准确率。其中,亲子鉴定准确率(%)表示程序鉴定出的真实错误系谱数占总错误系谱数的百分比。亲本推断准确率(%)表示,在进行亲本推断时,所推断出的疑似亲本为真实亲本的后代个体数,占真实错误系谱中所有后代个体的百分比。运行效率主要以程序运行时间来表示,分为亲子鉴定和亲本推断两部分时间。亲子鉴定时间表示,鉴定待测系谱正确与否所需的时间。亲本推断时间表示,对错误系谱进行亲本推断所需的时间。以上每种设计都进行10次重复计算,以消除随机误差的影响。

2 结果与分析

2.1 基因型数据

经过质控,600 000个标记剩余575 299个SNP位点用于后续研究。全部SNPs的MAF呈均匀分布(图2),平均MAF为0.28。

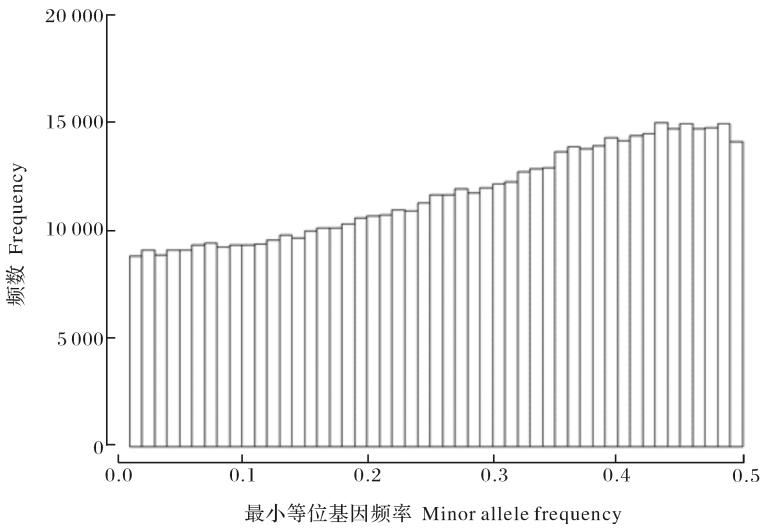

图2 最小等位基因频率分布

Fig.2 Histogram of minor allele frequency

2.2 判定结果准确率的影响因素

1)标记数对该方法准确率的影响。在不同系谱错误率下,利用不同标记数进行亲子鉴定和亲本推断的结果表明,无论标记数是多少,亲子鉴定准确率

均达到了 100%。然而，SNP 标记数为 100、200 时，平均推断准确率均不能达到 100%，分别为 94%、99%。但当 SNP 标记数为 300 以上进行推断时，推断准确率都达到了 100%（表 2）。

表 2　不同标记数下亲本推断的准确率

Table 2　Accuracy of paternity inference with variable number of markers　%

标记数/个 Number of markers	系谱错误率/% Pedigree error rate					
	5	10	15	20	25	30
100	94	96	92	97	94	91
200	98	100	99.3	98	99.8	99.7
300	100	100	100	100	100	100
>500	100	100	100	100	100	100

2）SNP 标记 MAF 对该方法准确率的影响。在不同 SNP 标记 MAF 下，测试了该方法在不同系谱错误率下进行亲子鉴定和亲本推断的准确率，结果发现，当筛选 MAF 在[0.01，0.2）内的标记进行亲子鉴定和亲本推断时，亲子鉴定准确率能达到 100%，但推断准确率只能达到 95%。而在[0.2，0.5]区间内时，不同系谱错误率下的鉴定和推断准确率均达到 100%（表 3）。

表 3　SNP 标记最小等位基因频率对亲子鉴定及亲本推断准确率的影响

Table 3　Accuracy of parentage identification and paternity inference using SNP markers with different minor allele frequency (MAF)

%

最小等位基因频率 MAF		系谱错误率/% Pedigree error rate					
		5	10	15	20	25	30
0.01～0.2	亲子鉴定 Parentage identification	100	100	100	100	100	100
	亲本推断 Paternity inference	96	94	94	94	97	93
0.2～0.5	亲子鉴定 Parentage identification	100	100	100	100	100	100
	亲本推断 Paternity inference	100	100	100	100	100	100

2.3　运行效率

1）标记数对该方法运行效率的影响。在不同系谱错误率下，利用不同标记数进行亲子鉴定和亲本推断的结果表明，亲子鉴定时间随标记数的增加而增加，两者呈明显线性相关（表 4）。其中，平均每增加 10 000 个 SNP 标记，鉴定耗时量仅增加 6 s。当标记数增加到 50 000 时，鉴定时间仅为 30.7 s，并且准确率为 100%。推断时间也随标记数的增加而增加，两者也存在线性关系（表 5），但相关程度较标记数与鉴定时间之间的相关程度偏低，并且耗时也相对较多，当标记数增加到 50 000 时，推断时间需 109.57 s。在相同标记数时，鉴定时间并没有随着系谱错误率的增加呈线性增长（表 4）。而推断时间是随着系谱错误率的增加而增加，即呈线性相关（表 5）。

表 4　不同标记数在不同系谱错误率下的鉴定耗时量

Table 4　Time demanding for parentage identification with variable number of markers and different pedigree error rate　s

标记数/个 Number of markers	系谱错误率/% Pedigree error rate					
	5	10	15	20	25	30
100	0.29	0.27	0.28	0.30	0.29	0.28
500	0.50	0.49	0.48	0.49	0.49	0.47
1 000	0.70	0.71	0.72	0.72	0.72	0.71
5 000	2.54	2.50	2.51	2.52	2.51	2.51
10 000	5.13	5.13	5.13	5.14	5.13	5.14
15 000	7.91	7.87	7.89	7.87	7.87	7.80
20 000	10.83	10.88	10.86	10.94	10.93	10.99
25 000	14.67	14.66	14.48	14.60	14.68	14.51
50 000	30.29	30.64	30.86	31.07	30.52	30.77

表5 不同标记数在不同系谱错误率下的亲本推断耗时量

Table 5 Time demanding for paternity inference with variable number of markers and different pedigree error rate s

标记数/个 Number of markers	系谱错误率/% Pedigree error rate					
	5	10	15	20	25	30
100	0.48	0.84	1.18	1.71	1.96	2.45
500	1.18	2.11	3.02	4.34	5.13	6.37
1 000	1.95	3.59	5.14	7.36	8.68	10.96
5 000	8.23	15.16	21.73	31.23	36.75	46.67
10 000	16.08	29.65	42.62	61.30	71.97	91.26
15 000	24.68	45.62	65.42	94.18	111.46	140.58
20 000	35.24	64.50	93.71	135.85	160.97	202.92
25 000	47.01	87.44	126.87	185.06	218.20	277.05
50 000	109.57	208.78	302.01	440.57	526.21	657.83

2)系谱数对该方法运行效率的影响。在标记数为50 000下,测试了该方法在不同系谱数下进行亲子鉴定的耗时量。测试结果表明,亲子鉴定时间和系谱数也呈明显的线性关系。其中,系谱数平均每增加100条,耗时增量为3.23 s,完成1 000条系谱的鉴定时间仅用31.47 s,并且准确率为100%(表6)。

表6 不同系谱数的鉴定耗时量

Table 6 Time demanding for parentage identification with different size of pedigrees s

标记数/个 Number of markers	系谱数 Pedigree size									
	100	200	300	400	500	600	700	800	900	1 000
50 000	2.62	5.42	8.18	11.10	14.07	17.46	20.90	25.69	27.28	31.47

3 讨 论

本研究提出了一种基于双亲全基因组SNP标记进行亲子鉴定及亲本推断的方法,可在畜禽群体中实现利用双亲高密度SNP标记信息对系谱进行检验并校正,通过一系列模拟研究对方法准确率的影响因素及运行效率进行详细研究。

目前,用于亲子鉴定的方法很多,但这些方法的适应范围、准确性和运行效率不同。本研究所提出的方法对标记数量的适用范围较为广泛,从少量(几百个)到海量SNP标记信息均可。而与之相比,部分其他方法计算复杂,对标记数量有一定约束,如Cervus[26]。同时,本研究的方法适用于SNP标记,这也是目前畜禽群体遗传研究最常用的一种标记类型。

对该方法运行准确性进行分析可知,在标记数>100时即可达到亲子鉴定100%准确,标记数>300时可达到亲本推断100%准确(表2),这表明该方法非常适用于当前高密度SNP芯片群体数据分析。在标记数量较少时,如少于100个标记,建议使用其他方法或软件完成亲子鉴定工作。除标记数外,我们也对MAF对该方法准确性的影响进行了研究,结果表明:当MAF<0.2时,亲子鉴定准确率能达到100%,而亲本推断准确率则略有降低,说明低MAF会影响亲本推断准确性。因此,在使用该方法时若要保证准确性,需考虑标记数量及MAF分布。

在运行效率方面,该方法也能够快速满足现有大量的研究需求,如对群体规模1 000、标记数量50 000的群体,用该方法进行亲子鉴定所需时间也在数十秒以内(表4)。同时,最大内存使用量约为5.8 GB,而将SNP数降低到25 000时,内存使用量约为3.3 GB。基于标记数对推断准确性的影响验证结果(表2),300个SNP标记即可准确地完成亲子鉴定,而此时普通PC可完全满足计算的内存需求。对运行效率的影响因素的分析结果表明:标记数、系谱错误率和系谱数都是影响该方法运行效率的关键因素。标记数与运行效率呈线性关系(表4)。系谱错误率对该方法的鉴定效率并无明显影响,只对推断效率有明显影响,且呈线性相关的关系(表5)。而系谱数对鉴定效率也成线性关系(表6)。

在对运行效率进行分析时,我们发现该方法在完成亲子鉴定以后的亲本推断过程占用了整个程序运行的大部分时间(表4和表5),因此,对于无亲本推断需求的研究者来说,关闭亲本推断功能,可进一

步提高方法的运行效率。亲本推断较为耗时的原因是：亲本推断时，在无其他先验信息的情况下，该方法需在整个群体中对个体的所有可能亲本进行检测。

综上，本研究基于孟德尔遗传原理，设计并开发了基于双亲及后代基因型的亲子鉴定及亲本推断程序。该方法能够利用双亲高密度 SNP 标记进行亲子鉴定及亲本推断。模拟研究结果表明：该方法在标记数＞100 时即可准确进行亲子鉴定，标记数＞300 即可准确进行亲本推断。标记多态性较低时会降低方法的准确性。标记数、系谱数对亲子鉴定和亲本推断的运行时间均为线性关系，而系谱错误率仅影响亲本推断速度。因此，在当前基因组大数据应用环境下，该方法值得进一步推广应用以提高群体基因组学相关研究的效率。

参　考　文　献

[1] BANOS G，WIGGANS G R，POWELL R L.Impact of paternity errors in cow identification on genetic evaluations and international comparisons[J]. Journal of dairy science，2001，84(11)：2523-2529.

[2] 汪湛，田雨泽，刘和凤.应用血型分析技术对奶牛亲子关系正确率的调查初报[J].中国畜牧兽医，2005，32(3)：22-23.

[3] 郭刚，周磊，刘林，等.利用 SNP 标记进行北京地区中国荷斯坦牛亲子推断的研究[J].畜牧兽医学报，2012(1)：44-49.

[4] 韩春梅，张嘉保，高庆华，等.微卫星 DNA 在吉戎兔亲子鉴定中的应用研究 [J].遗传，2005，27(6)：903-907.

[5] ØYSTEIN S，GLOVER K A，BARLAUP B T，et al.Microsatellite DNA used for parentage identification of partly digested Atlantic salmon (*Salmo salar*) juveniles through non-destructive diet sampling in salmonids[J]. Marine biology research，2014，10(3)：323-328.

[6] YU G C，TANG Q Z，LONG K R，et al.Effectiveness of microsatellite and single nucleotide polymorphism markers for parentage analysis in European domestic pigs[J].Genetics & molecular research，2015，14(1)：1362-1370.

[7] FERRIE G M，COHENO R，SCHUTZ P，et al.Identifying parentage using molecular markers：improving accuracy of studbook records for a captive flock of marabou storks (*Leptoptilus crumeniferus*)[J].Zoo biology，2013，32(5)：556-564.

[8] SANDERS K，BENNEWITZ J，KALM E.Wrong and missing sire information affects genetic gain in the Angeln dairy cattle population[J].Journal of dairy science，2006，89(1)：315-321.

[9] RENDEL J.Relationships between blood groups and the fat percentage of the milk in cattle[J].Nature，1961，189(1)：408-409.

[10] STORMONT C.Contribution of blood typing to dairy science progress[J].Journal of dairy science，1967，50(2)：253-260.

[11] 周磊，刘林，初芹，等.奶牛亲子鉴定应用的标记和方法研究进展[J].中国奶牛，2011(2)：26-29.

[12] JEFFREYS A J，BROOKFIELD J F，SEMEONOFF R.Positive identification of an immigration test-case using human DNA fingerprints[J].Nature，1985，317(6040)：818-819.

[13] KASHI Y，LIPKIN E，DARVASI A，et al.Parentage identification in the bovine using "deoxyribonucleic acid fingerprints"[J].Journal of dairy science，1990，73(11)：3306-3311.

[14] ALFORD R L，HAMMOND H A，COTO I，et al.Rapid and efficient resolution of parentage by amplification of short tandem repeats[J].American journal of human genetics，1994，55(1)：190-195.

[15] GLOWATZKI-MULLIS M L，GAILLARD C，WIGGER G，et al.Microsatellite-based parentage control in cattle[J]. Animal genetics，1995，26(1)：7-12.

[16] HEATON M P，HARHAY G P，BENNETT G L，et al.Selection and use of SNP markers for animal identification and paternity analysis in US beef cattle[J]. Mammalian genome，2002，13(5)：272-281.

[17] 张哲，罗元宇，李晴晴，等.一种基于高密度遗传标记的亲子鉴定方法及其应用 [J].遗传，2014，36(8)：835-841.

[18] LEE S Y，CHO G J.Parentage testing of thoroughbred horse in Korea using microsatellite DNA typing[J].Journal of veterinary science，2006，7(1)：63-67.

[19] RON M，BLANC Y，BAND M，et al.Misidentification rate in the Israeli dairy cattle population and its implications for genetic improvement[J].Journal of dairy science，1996，79(4)：676-681.

[20] 初芹，张毅，孙东晓，等.应用微卫星 DNA 标记分析荷斯坦母牛系谱可靠性及影响因素[J].畜牧兽医学报，2011，42(2)：163-168.

[21] 郭立平，徐丽，朱淼，等.西门塔尔牛微卫星亲子鉴定体系的优化[J].畜牧兽医学报，2013，44(6)：871-879.

[22] WERNER F A O，DURSTEWITZ G，HABERMANN F A，et al.Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds[J].Animal genetics，2004，35(1)：44-49.

[23] GOODNIGHT K F，QUELLER D C.Computer software for performing likelihood tests of pedigree relationship using genetic markers[J].Molecular ecology notes，1999，8(7)：1231-1234.

[24] DUCHESNE P，GODBOUT M H，BERNATCHEZ L.PAPA (package for the analysis of parental allocation)：a computer program for simulated and real parental allocation[J].Molecular ecology notes，2002，2(2)：191-193.

[25] GERBER S，CHABRIER P，KREMER A.FAMOZ：a software for parentage analysis using dominant，codominant and uniparentally inherited markers[J].Molecular ecology notes，2003，3(3)：479-481.

[26] KALINOWSKI S T，TAPER M L，MARSHALL T C.Revising how the computer program CERVUS accommodates genoty-

ping error increases success in paternity assignment[J].Molecular ecology,2007,16(5):1099-1106.
[27] CERVINO A C,LI G,EDVARDS S,et al.Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels[J]. Genomics,2005,86(5):505-517.
[28] XU Z,TAYLOR J A.SNPinfo:integrating GWAS and candidate gene information into functional SNP selection for genetic association studies[J].Nucleic acids research,2009,37(2):600-605.
[29] LEGARRA A,AGUILAR I,MISZTAL I.A relationship matrix including full pedigree and genomic information[J].Journal of dairy science,2009,92(9):4656-4663.
[30] ZHANG Z,LI X,DING X,et al.GPOPSIM:a simulation tool for whole-genome genetic data[J].BMC genetics,2015,16(1):1-6.

Parentage identification and paternity inference based on SNP markers in livestock population

LUO Yuanyu WU Peng HE Jinlong CHEN Zanmou ZHANG Hao LI Jiaqi ZHANG Zhe

College of Animal Science,South China Agricultural University/Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding/National Engineering Research Center for Breeding Swine Industry,Guangzhou 510642,*China*

Abstract A computer program was developed for parentage identification and paternity inference based on the single nucleotide polymorphisms (SNPs) of parents and offspring from known livestock population.Furthermore,the approach was tested with a simulated population with 600 000 SNP markers.Results showed that at least 100 SNPs are needed for correct parentage identification and a minimum of 300 SNPs for correct paternity inference.However,using markers with low average minor allele frequency can decrease the accuracy of paternity inference.The time for parentage identification of 1 000 pedigrees with 10% errors genotyped with 50 000 SNPs was 332.87 s,which showed linearly relationship with the number of markers and individuals.In this study,a computer program was developed for parentage identification and paternity inference according to Mendel's law and testified with markers from known genotypes of parents and their offspring.The program runs fast and simply with higher accuracy,and hence can be implemented potentially in relevant studies of genomics in livestock population.

Keywords parentage identification; paternity inference; SNP marker; Mendelian error; accuracy

(责任编辑:边书京)