



高密度单核苷酸多态性的系谱推断效能研究*

李 晶^{1,3)**} 孙一杰^{2,3)**} 赵雯婷³⁾ 汤子琛⁴⁾ 刘 京^{3)***} 李彩霞^{1,2,3)***}⁽¹⁾ 河北医科大学法医学院, 河北省法医学重点实验室, 河北省法医分子鉴定协同创新中心, 石家庄 050017;⁽²⁾ 中国人民公安大学侦查学院, 北京 100038;⁽³⁾ 公安部鉴定中心, 法医遗传学公安部重点实验室, 北京市现场物证检验工程技术研究中心, 北京 100038;⁽⁴⁾ 江苏师范大学生命科学学院, 江苏省系统发育与比较基因组学重点实验室, 徐州 221000

摘要 目的 研究不同量级单核苷酸多态性 (single-nucleotide polymorphism, SNP) 位点组合, 基于筛选 SNP 位点集进一步提高远亲缘关系的预测能力。**方法** 首先选取 3 种芯片中国基因分型芯片 (Chinese genotyping array, CGA, Illumina)、全球筛查芯片 (global screening array, GSA, Illumina)、23 魔方 V2 版高密度 SNP 芯片 (23MF_V2 high-density SNP array, Affy, Thermo Fisher Scientific (formerly Affymetrix)) 位点进行合并、质控, 筛选得到一组高密度 SNP 位点集 (1 180 k); 从 161 份全基因组测序数据中获取 1 180 k 位点集, 使用共祖片段 (identity-by-descent, IBD) 算法进行亲缘关系推断, 通过 IBD 片段长度和预测准确性的变化趋势评估该位点集的亲缘关系推断能力。**结果** 经过筛选后, 得到 1 184 334 个常染色体 SNP 位点集 (下文称高密度 SNP 位点集)。与 3 种芯片位点集的平均结果相比较, 高密度 SNP 位点集增加了总 IBD 片段长度以及 1~9 级的平均 IBD 片段长度; 8 级置信区间准确率为 70.97%, 提高了 3.50%; 1~8 级平均置信区间准确率为 91.39%, 提高了 1.00%; 8、9 级假阴性率分别降低 2.42%、6.76%。高密度 SNP 位点集的 1~8 级亲缘关系推断系统效能达 98.91%。通过随机减少位点结果发现, 增加 SNP 位点数量能提升较远亲缘关系推断能力。**结论** 高密度 SNP 位点集显著增强远缘关系推断效能, 可精准覆盖 1~8 级亲缘关系, 且 1~8 级平均置信区间准确率稳定在 90% 以上。本研究发现, SNP 位点数量增加, 可以提高远亲缘预测能力。

关键词 高密度单核苷酸多态性位点, 法医单核苷酸多态性系谱推断技术, 全基因组测序技术, 亲缘关系推断, 共祖片段算法

中图分类号 RD89, D919

DOI: 10.3724/j.pibb.2025.0340

CSTR: 32369.14.pibb.20250340

法医 DNA 领域常用的遗传标记短串联重复序列 (short tandem repeat, STR) 仅适用于 2 级以内的近亲缘关系的分析^[1]。单核苷酸多态性 (single nucleotide polymorphism, SNP) 被认为是第三代 DNA 遗传标记, 具有在基因组中分布密度高、突变率低等显著优势。近年来, SNP 被广泛用于法医 DNA 领域, 尤其在亲缘关系鉴定中发挥着重要的作用^[2-3]。2018 年, 美国著名的“金州杀手案”破获, 让法医 SNP 系谱推断技术 (forensic SNP genealogy inference, FSG) 受到广泛关注^[4-5], 并被《科学》(Science) 杂志评选为 2018 年十大科学突破之一。法医 SNP 系谱推断技术是基于高密度 SNP 分型数据, 分析案件检材与特定样本之间或与社会公共 DNA 数据库中个体之间的亲缘关系等级, 进而为案件侦查提供线索^[6-7]。目前, 国内外警方

应用该技术已成功破获多起案件^[5, 8-10]。例如: 瑞典警方利用 FSG 技术在 GEDmatch、FamilyTree 等数据库比中远亲缘关系个体, 成功破获了 16 年前的杀人案^[11] 和 15 年前的双重谋杀案^[12]; 公安部鉴定中心刘京等^[13-15] 利用 FSG 技术比中近亲缘破获古汉墓被盗案^[13]、比对出 7 级亲属远亲缘关系个体, 成功破获 14 年积压冷案^[14] 和 23 年积压冷案^[15]; 吉林大学古 DNA 实验室运用法医 SNP 系谱

* 公安部鉴定中心基本科研业务费专项资金 (2024JB026, 2024JB043), 国家自然科学基金 (82171870) 和国家重点研发计划 (2022YFC3341004) 资助。

** 并列第一作者。

*** 通讯联系人。

李彩霞 Tel: 010-83752706, E-mail: 90650101@hebmu.edu.cn

刘京 Tel: 18211057899, E-mail: biojing@yeah.net

收稿日期: 2025-07-18, 接受日期: 2026-01-04

推断技术, 鉴定出 20 年前的案件遗骨与其父母的亲子关系^[16]。

高密度 SNP 分型数据需要通过 SNP 芯片和全基因组测序 (whole genome sequence, WGS) 等技术获得^[17-18]。其中 SNP 芯片是目前常用的 SNP 分型技术, 商业性的 SNP 芯片包括: 中国基因分型芯片 (Chinese genotyping array, CGA, Illumina)、全球筛查芯片 (global screening Array, GSA, Illumina)、23 魔方 V2 版高密度 SNP 芯片 (23MF_V2 High-Density SNP Array, Affy, Thermo Fisher Scientific (formerly Affymetrix)) 等, SNP 位点数量在 50 万~80 万之间。共祖片段 (identity-by-descent, IBD) 算法^[19]是目前国内外常用的系谱推断算法^[20-21], 是基于高密度 SNP 数据, 计算两样本间 IBD 片段的总厘摩 (centimorgan, cM) 长度进而分析样本间的亲缘关系。在遗传学上, 有亲缘关系的个体之间会共享其共同祖先的 DNA, 亲缘关系越近, 共享 DNA 片段越长^[22]。研究表明, 基于 IBD 算法对 SNP 芯片数据进行分析能推断 7 级以内的亲缘关系^[23]。

由于 SNP 芯片的位点和数量固定, 且位点通常都在百万级以下, 即使使用高检出率 ($\geq 99\%$) 的芯片数据, 亲缘关系的预测值和真实值之间通常存在 1~3 级的误差, 并且亲缘关系越远误差越大, 约 10% 的 3 代表亲 (7 级) 和 54% 的 4 代表亲 (9 级) 的 IBD 片段未被检测到^[2]。SNP 芯片技术对 DNA 质量要求很高, DNA 投入量需要 200 ng, 并对微量、降解 DNA 的检测能力不足^[24], 制约了系谱推断技术的法医学应用潜力^[25-27]。鉴于法医样本经常因暴露于各种环境伤害而受到损害、降解, 且数量和质量可能显著降低, 有必要采用更全面的 SNP 分析技术, 以生成足够的 SNP 数据用于法医系谱分析。相比之下, WGS 是微量、降解样本的首选测序方法^[12, 28], 能够获取样本的全部 SNP 分型数据, 且数据测序覆盖深度可达 $30\times\sim 100\times$ 。例如, 瑞典警方通过提取 2003 年的骨 DNA 进行 WGS, 获取 100 万个 SNP 基因分型, 在 GEDmatch 中比中候选亲属列表, 为案件侦破提供线索^[11]。

因此, 本研究基于国内真实家系的 WGS 数据, 旨在探究采用更高密度的 SNP 位点集能否提高 IBD 算法准确性, 进而提升远亲缘预测能力, 为提升法医 SNP 系谱推断技术的法医学应用潜能提供有力的技术支持。

1 材料与方法

1.1 样本来源

本研究采集 3 个志愿者家系共 161 份汉族唾液样本^[29], 其中包含 1 321 对 1~9 级亲缘关系, 其中包含 104 对 1 级 (1st)、88 对 2 级 (2nd)、138 对 3 级 (3rd)、182 对 4 级 (4th)、240 对 5 级 (5th)、187 对 6 级 (6th)、184 对 7 级 (7th)、124 对 8 级 (8th)、74 对 9 级 (9th) 亲缘关系, 11 559 对无亲缘关系 (表 1)。本研究所有参与者均签署知情同意书, 并通过了公安部物证鉴定中心伦理委员会批准 (编号: 2023-004)。

Table 1 The distribution of kinship degree in pedigree samples

Survey	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	Sum
degree										
Number	104	88	138	182	240	187	184	124	74	1 321

1.2 DNA提取与检测

使用 MagAttract M48 DNA Manual Kit (Qiagen, Hilden, Germany) 提取唾液样本 DNA。使用 NanoDrop 2000c 超微量分光光度计 (Thermo Fisher Scientific, Waltham, USA) 检测 DNA 浓度, DNA 投入体积 40 μ l 且 DNA 浓度达到 5 μ g/L 以上时, 即可采用 MGIEasy 通用 DNA 制备试剂套装 (MGI-Shenzhen, China) 进行文库制备, 取适量基因组 DNA 进行片段化, 进行磁珠筛选, 筛选合适片段后, 配制反应体系进行 DNA 末端修复及 3' 端 A 碱基添加。随后, 配制接头连接体系, 使接头与 DNA 片段连接形成文库。文库经 PCR 扩增后, 进行质检。合格文库经变性、环化反应, 消化未环化线性 DNA, 形成单链环状 DNA。再经 phi29 滚环复制, 生成含多拷贝的 DNA 纳米球 (DNA nanoball, DNB), 通过高密度 DNA 纳米芯片技术置于芯片网状小孔, 于 DNBSEQ-G400 (MGI-Shenzhen, China) 平台对 161 份 DNA 进行 PE150 的 WGS。

1.3 数据分析

使用 RTyper 21 试剂盒 (新海生物, 苏州, 中国) 对所有 DNA 样本进行 STR 分型, 包括 20 个常染色体位点和 1 个性染色体位点。扩增后 PCR 产物使用 3500 基因分析仪 (Thermo Fisher Scientific, Massachusetts, America) 和 GeneMapper ID-X v1.5 软件进行分析, 通过 STR 分型结果验证亲子关系。

全基因组数据分析流程如下：通过 fastp 软件对下机数据 fastq 文件进行质量控制和过滤，去除低质量片段和接头序列；使用 bwa-mem2 软件比对到人类参考基因组上 (hg19)，并按照染色体位置进行排序；使用 Genome Analysis Toolkit 的 MarkDuplicates 工具剔除 PCR 重复序列，再用 BQSR 功能校正碱基质量^[30]。随后使用 Genome Analysis Toolkit HaplotypeCaller 对每个样本分别进行变异位点识别，生成包含所有位点（包括纯合参考）的基因组 VCF (gVCF) 文件，精准提取目标 SNP^[30]。接着依据测序深度>5、基因型质量值>10 的标准过滤低质量位点，并将 VCF 文件转换为 plink 二进制数据格式。

1.4 高密度 SNP 位点集

本研究选用 3 种商业化芯片的 SNP 位点集：CGA 芯片包含 71 万个 SNP、GSA 芯片包含 70 万个 SNP、Affy 芯片包含 54 万个 SNP。首先对 3 种芯片位点集进行合并，共 150 万 SNP，随后保留与千人基因组东亚人群数据集中共有的 122 万个 SNP，最后去除性染色体和线粒体 SNP 位点，剩余 118 万个常染色体 SNP 作为高密度 SNP 位点集，用于后续分析。

1.5 亲缘关系推断分析

将 1.3 部分最后获取的数据，使用 IBD 算法计算两两样本间的 IBD 长度进行亲缘关系推断。IBD 算法流程是以 Plink、VcfTools 软件为辅助，通过 Eagle 软件对样本数据进行同源染色体分离，然后使用 Germline 软件识别所有样本对之间相同的 DNA 片段（原始错配参数值为 $\text{err_het}=3$ 、 $\text{err_hom}=2$ ，缩写为 err_het_3_hom_2 ），最后通过 Ersa 软件预测个体间亲缘关系。其中 Germline 的错配参数值 (err_hom 、 err_het) 是指允许错误匹配的纯合和杂合遗传标记的最大数量^[31]。

对不同位点集预测的 IBD 长度，计算平均值 ($\pm$ 方差)。而后，针对不同位点集每一等级下的平均 IBD，评估其检出能力。通过绝对准确率 (accuracy, AC)、置信区间准确率 (confidence interval accuracy, CIA)、假阴性率 (false negative, FN)、假阳性率 (false positive, FP) 和预测可信度 (prediction credibility, PC) 指标评估预测准确性。其中 AC 是指预测亲缘关系等级在调查亲缘关系（真实亲缘关系）等级关系对数占该等级调查关系对数的比重，CIA 是指预测亲缘关系等

级在调查亲缘关系等级的上下一级浮动的关系对数占该等级调查关系对数的比重，FN 是指调查存在亲缘关系但是被预测为无关的对数占该等级调查关系对数的比重，FP 是指无关被预测为有关的对数占所有调查关系对数的比重，PC 是指在被预测为该等级的全部的亲缘关系对数中，经实际调查关系对数所占的比重。

1.6 IBD 长度阈值筛选

IBD 长度阈值是 IBD 算法检测两两个体间 IBD 长度的最低限。本研究设置 0、1、2、3、4、5、6、7、8、9 cM 共 10 个阈值，评估不同 IBD 片段长度阈值对预测准确性的影响^[23]。

1.7 SNP 位点数量对亲缘关系推断能力的影响

通过每一等级下平均 IBD 长度以及预测准确性参数评估亲缘关系推断能力。使用 161 份真实家系样本 DNA 的全基因组测序数据（平均深度 30 $\times$ ），提取所需的数量的 SNP 位点集进行该研究，对比高密度 SNP 位点集与如下位点集的亲缘关系推断能力。

1.7.1 芯片 SNP 位点集

比较高密度 SNP 位点集与 3 种芯片 SNP 位点集以及 3 种芯片位点交集的亲缘关系推断能力，评估高密度 SNP 位点集相较于芯片位点的亲缘关系推断能力。

1.7.2 随机减少位点集

使用 R 语言编写脚本对高密度 SNP 位点集进行随机位点删减，研究不同数量梯度的位点对亲缘关系推断能力的影响。本研究将随机删减位点数量范围设置为 110 万~10 万，每递减 10 万为 1 组共 11 组，每组位点随机筛选 10 次，提取 161 份 WGS 数据中的相应位点集，进行亲缘关系推断，每组推断结果取 10 次平均值。评估不同数量的位点集对亲缘关系推断能力的影响。

1.8 IBD 推断参数筛选

为进一步提升 118 万个 SNP 的高密度位点集的远亲缘关系预测能力，本研究通过调整 IBD 算法中 Germline 的错配参数值 (err_hom 、 err_het)，以提升样本对间共享 IBD 的检出能力，进而提高亲缘关系的预测准确性。将 err_hom 值设定在 2~19 的范围，限制 $\text{err_het}=\text{err_hom}+1$ 。同时考虑假阳性对亲缘预测准确性的影响，选取假阳性率值在 10% 以内的错配参数结果进行筛选。

2 结 果

2.1 全基因组测序质量评估

图 1 显示高密度 SNP 位点集均匀地分布于 22 对常染色体上, 符合高密度 SNP 位点在人类基因组中均匀分布的标准。利用 DNBSEQ-G400 对 161

份样本开展 WGS, 结果显示 (表 S1): 样本的平均比对率为 98.78%, 平均测序深度为 36.95×。就覆盖度数据而言, 1×覆盖度和 4×覆盖度的平均值均大于 98%, 10×覆盖度的平均值是 97.12%, 20×覆盖度的平均值是 89.97%。所有样本的比对率、测序深度和覆盖度等参数详见表 S1。

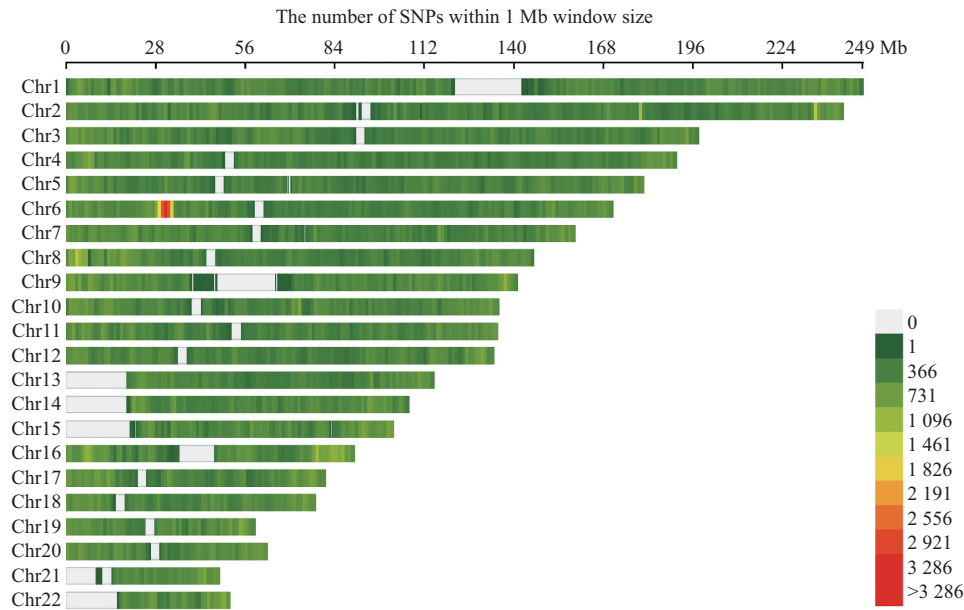


Fig. 1 Distribution of the high-density SNP panel on autosomes

2.2 IBD片段长度阈值筛选

前期探索表明不同位点集可能存在不同最优 IBD 长度阈值 [17, 23-24]。为减少不同最优 IBD 长度阈值这一对后续研究的影响, 首先对 IBD 长度阈值进行筛选。从 161 份 WGS 家系数据中提取高密度 SNP 位点集数据, 设置 0、1、2、3、4、5、6、7、8、9 共 10 个不同长度阈值 (cM) 分别进行 IBD 长度分析, 通过不同 cM 阈值下的预测准确性筛选最优 IBD 片段长度 [23]。

图 2 展示了不同 cM 阈值的预测准确性结果 (表 S2), 表 S2 和图 2 的 a、b、c 分别呈现了 AC、CIA、FN、FP 的变化情况, 其中 0、1、2 cM 的 IBD 长度预测结果相同。在表 S2 和图 2a 中, AC 和 CIA 值随 cM 阈值的增加而降低; 在图 2b 中, FN 值从 5 级亲缘开始出现, 随 cM 阈值的增加而升高; 在图 2c 中, FP 值随 cM 阈值的增加而降低, 但变化幅度较小, 最大相差 0.16%。综上所述, 预测准确性随 cM 阈值增加而降低, 0、1、2 cM 的预测准确性最高。本研究进一步通过 Wilcoxon 秩和检验

方法, 将 0、1、2 cM 结果与其他阈值结果的 CIA 值及 FN 值进行对比分析, 以评估两组数据之间是否具有显著性差异。在 CIA 值上, 0、1、2 cM 与 3 cM 相比无显著性差异 ($P>0.05$), 与 4~9 cM 相比有显著性差异 ($P<0.05$), 表明 0、1、2 cM 相较于 4~9 cM 在 CIA 值上有显著提升; 在 FN 值上, 0、1、2 cM 与 3~9 cM 相比无显著性差异 ($P>0.05$), 表明不同 cM 阈值对 FN 值无显著影响; 在 FP 值上, 0、1、2 cM 与 3~9 cM 相比无显著性差异 ($P>0.05$)。与 3 cM 预测结果相比, 0、1、2 cM 的 AC 提升幅度较小。不过 0、1、2 cM 的 CIA 值在 5 级、9 级分别提升 0.41%、1.36%, FN 值在 9 级降低 1.35%, 而 FP 值在 9 级仅相差 0.02%。综上所述 0、1、2 cM 为最优 IBD 长度阈值。为避免由于低 IBD 长度阈值导致的冗余信息, 本研究选用 2 cM IBD 长度阈值进行以下研究。

2.3 SNP位点数量对亲缘关系推断能力的影响

2.3.1 芯片SNP位点集

将 118 万个常染色体 SNP 的高密度位点集标记

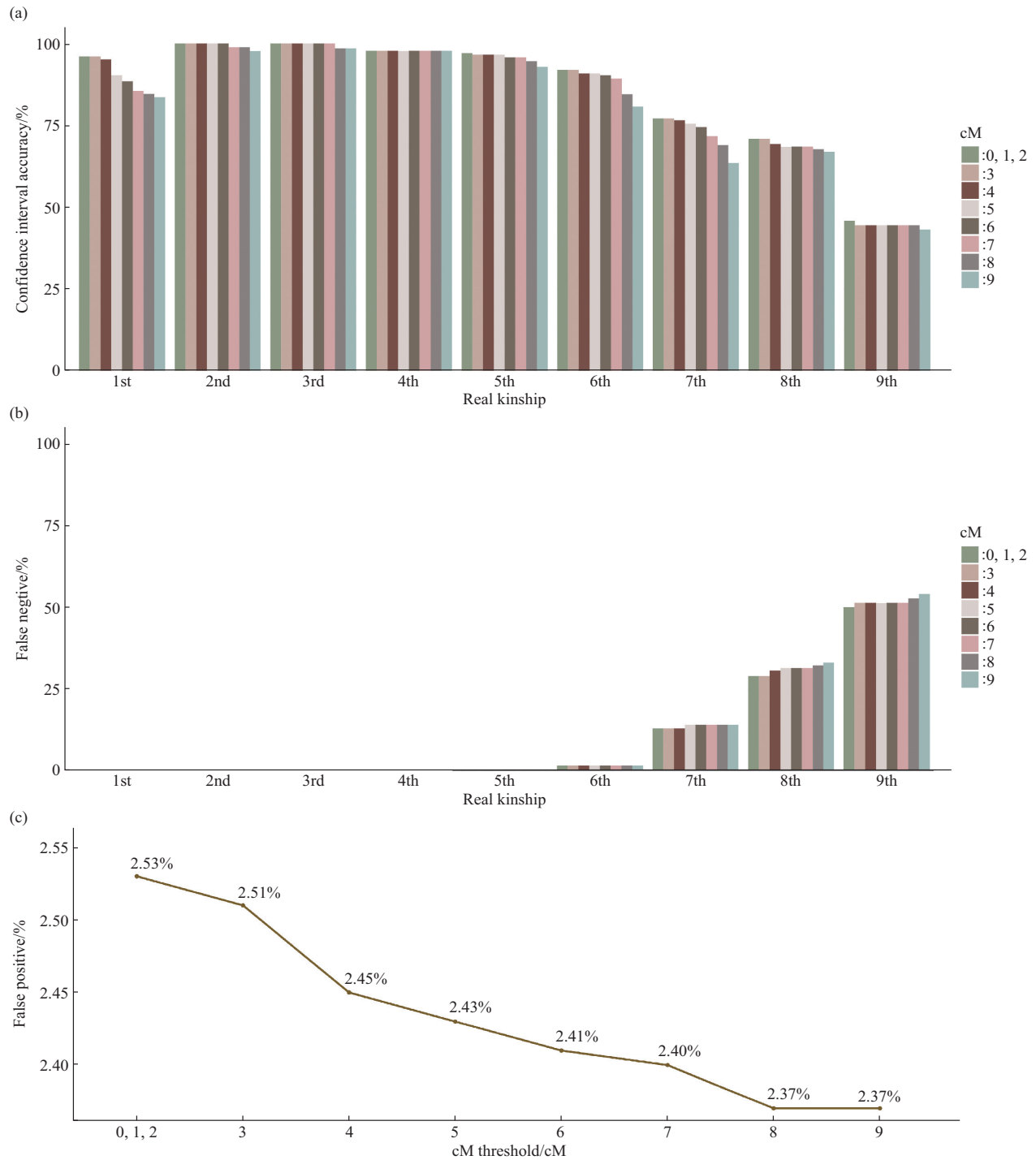


Fig. 2 IBD prediction accuracy results with different cM thresholds

Confidence interval accuracy (CIA): the probability that two related individuals are predicted to be within the range of 1st of their survey kinship; false negative (FN): the probability of predicting two related individuals as unrelated; false positive (FP): the probability of predicting two unrelated individuals as within 9th degree.

为P1、66万个常染色体SNP的CGA芯片位点集标记为P2、64万个常染色体SNP的GSA芯片位点集标记为P3、54万个常染色体SNP的Affy芯片位点集标记为P4、7万个常染色体的CGA、GSA、Affy

芯片的交集标记为P5。

a. IBD长度评估

通过对比不同位点集在不同等级下平均IBD长度的差异，分析P1对每一等级下平均IBD长度的

影响。表2为不同位点集在每一等级下的平均IBD长度, 结果以平均值(±方差)的形式呈现。通过对比发现, P1在每一等级下平均IBD长度都有提升。对比P1与P2、P3、P4、P5的平均IBD长度, 发现P1对平均IBD长度提升显著(秩和检验, $P<0.05$)。综上所述, 表明P1提高了IBD的检出能力, 为提高亲缘关系预测准确性提供了基础。

Table 2 The average IBD length of different SNP panel in each kinship degree (cM)

Locus set	P1	P2	P3	P4	P5
1st	3 134.05 (438.65)	3 088.48 (480.66)	3 126.27 (480.54)	3 122.12 (486.46)	2 352.55 (801.32)
2nd	1 699.11 (244.91)	1 641.6 (248)	1 657.41 (243.9)	1 646.21 (242.8)	1 116.47 (306.52)
3rd	833.58 (177.93)	815.86 (173.62)	822.71 (175.79)	818.13 (172.9)	520.99 (144.1)
4th	427.52 (118.48)	417.69 (116.43)	421.6 (116.81)	416.64 (116.45)	232.08 (88.88)
5th	204.11 (64.61)	199.18 (64.14)	201.31 (64.46)	197.66 (64.45)	99.71 (46.39)
6th	104.18 (47.18)	101.04 (47.09)	102.07 (47.62)	99.81 (46.72)	48.61 (34.59)
7th	51.37 (28.56)	48.98 (28.14)	49.38 (28.45)	49.5 (27.98)	21.21 (20.97)
8th	28.46 (20.32)	26.36 (19.64)	26.69 (19.64)	26.12 (19.45)	9.93 (14.28)
9th	16.03 (15.56)	14.39 (14.02)	14.49 (14.3)	14.59 (14.24)	4.38 (8.96)

Data are presented as mean (±S.D.); Wilcoxon test: (1) P1 vs P2, $P=2.56\times10^{-136}$; (2) P1 vs P3, $P=7.12\times10^{-71}$; (3) P1 vs P4, $P=6.7\times10^{-112}$; (4) P1 vs P5, $P=8.78\times10^{-92}$.

b. 不同位点集的预测准确性评估

图3为不同位点集的预测准确性结果(表S3, S4), 表S3、表S4和图3的a、b、c分别为PC、AC、CIA、FN、FP的变化情况。在表S4中, 与P2、P3、P4相比, P1的6~9级AC值平均提升1.96%、3.80%、1.88%、0.45%; 在图3a中, 相较于P2、P3、P4, P1的6~9级CIA值平均提升1.78%、2.36%、3.49%、4.95%; 与P5相比, P1的1~9级CIA值均呈现显著性提升(卡方检验, $P<0.05$)。在图3b中, 与P2、P3、P4相比, P1的8~9级FN值平均降低2.42%、6.76%; 并且与P5相比, P1的6~9级FN值均呈现显著性降低(卡方检验, $P<0.05$)。在图3c中, P1的FP值呈现显著性增加(卡方检验, $P<0.05$)。研究结果显示, 与P2、P3、P4、P5相比, P1显著提高了远亲缘预测准确性, 这与IBD长度的预测结果相一致。

2.3.2 随机减少位点集

为模拟现场生物检材的随机缺失位点的情形, 开展针对高密度SNP位点集进行随机减少位点研究。将随机减少位点集合设置为11组, 位点数量分别为110万、100万、90万、80万、70万、60万、50万、40万、30万、20万、10万, 将高密度SNP位点集的118万位点与随机减少位点集共同分析, 评估不同数量位点在全测序家系数据中对亲缘关系推断能力的影响。

a. IBD长度评估

图4为平均IBD长度在不同等级下的变化情

况。由图可知, 平均IBD长度随位点数量增加而增加。并且通过秩和检验分析相邻两组位点平均IBD长度的显著性差异, 发现位点数量每增加10万, 其平均IBD长度均存在显著性差异($P<0.05$)。结果显示, 随位点数量的增加, 总IBD长度呈现“先快后慢”的趋势。当位点数量小于30万时, 随着位点增加, IBD长度快速增长; 当位点数量处于30万和70万之间时, 随着位点增加, IBD长度缓慢增长; 当位点数增加到70万以上时, 变化开始趋于平缓。综上所述, IBD长度的检出能力随位点数量的增加而增加, 与芯片SNP位点集的IBD长度预测结果基本一致。

b. 随机减少位点对预测准确性的影响评估

图5为随机减少位点的预测准确性结果(表S5), 表S5和图5的a、b、c分别为AC、CIA、FN、FP的变化情况。在表S5和图5a中, AC和CIA值整体呈现随位点数量增加而增加; 在图5b中, 5级亲缘出现假阴性(平均为0.43%), 6~9级FN值随位点数量增加而降低; 在图5c中, FP值随位点数量增加呈现升高趋势, 最大相差2.04%。本研究结果显示, 增加SNP位点数量有助于提升预测准确性。进一步对比高密度SNP位点集与随机减少位点集在远亲缘关系6~9级CIA值、FN值以及FP值的显著性差异。结果显示: 高密度SNP位点集与20万以下位点集在7级CIA值上有显著性差异(卡方检验, $P<0.05$); 与10万位点集在7级FN值上有显著性差异(卡方检验, $P<0.05$); 与90万以及70

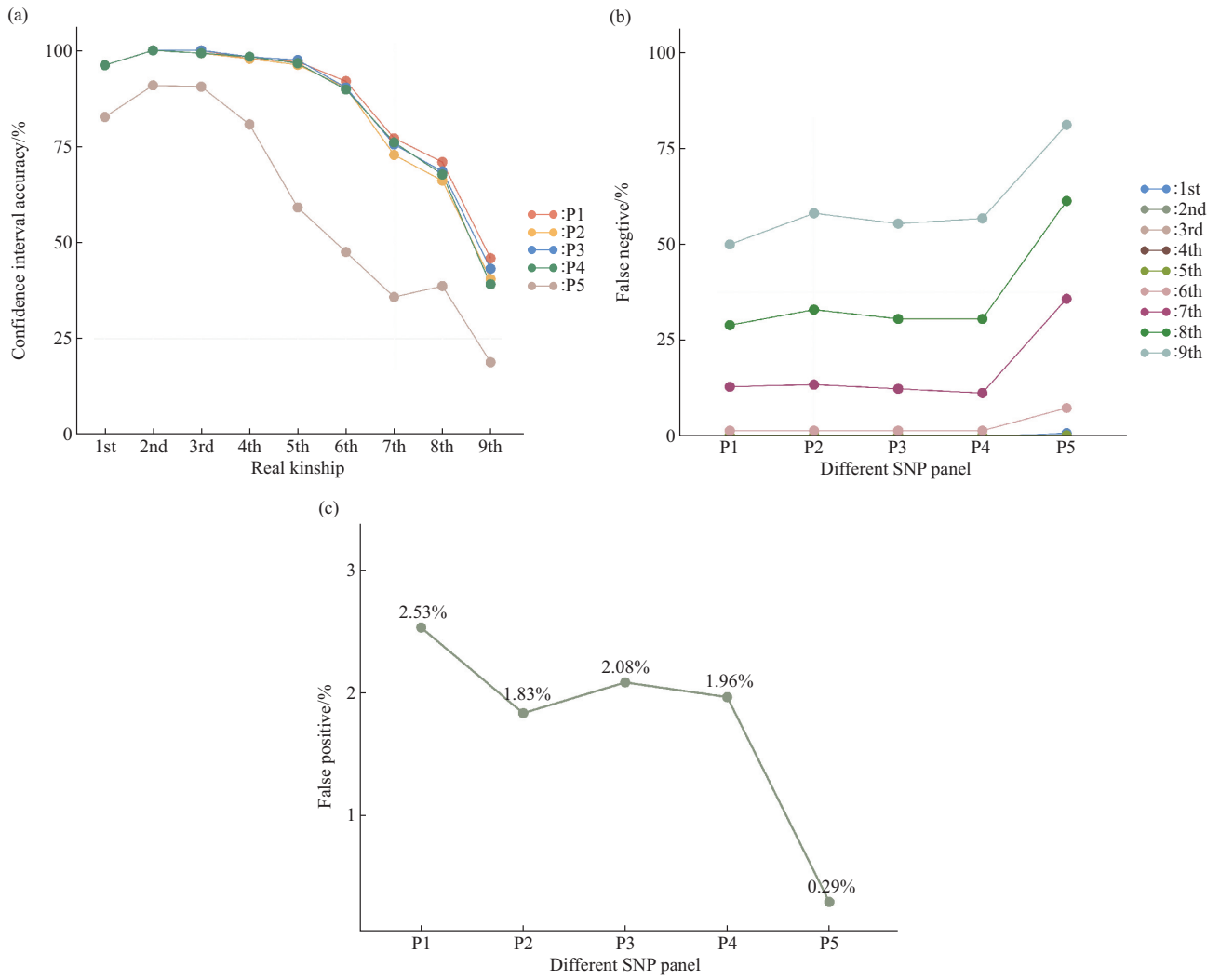


Fig. 3 Prediction accuracy results for different SNP panel

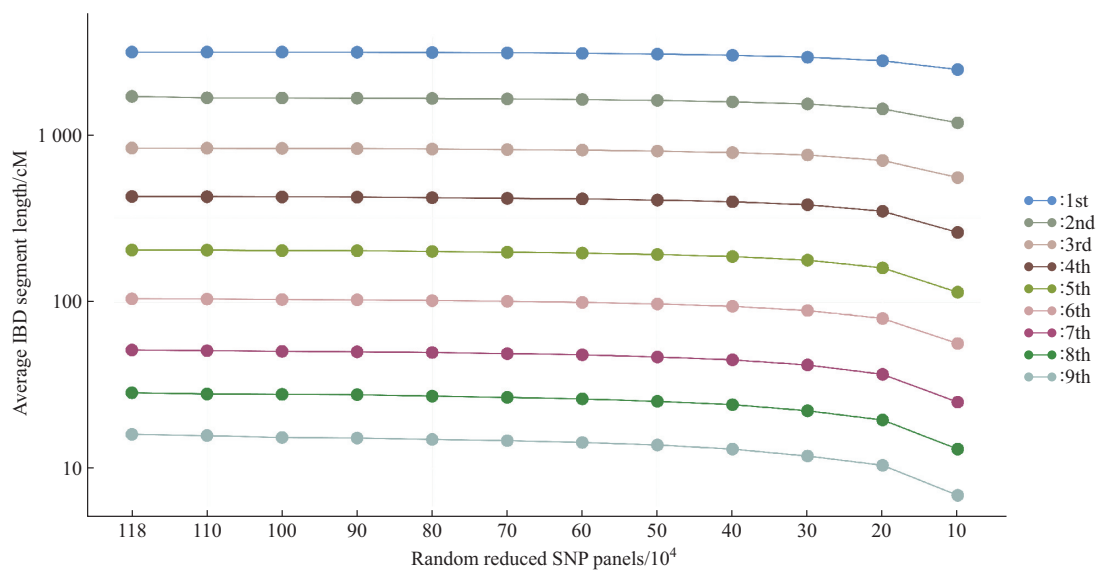


Fig. 4 The mean IBD length distribution of random reduced SNP panels

万以下位点集在 FP 值上有显著性差异(卡方检验, $P<0.05$)。结果表明, 位点数量增加对 FP 值的变化

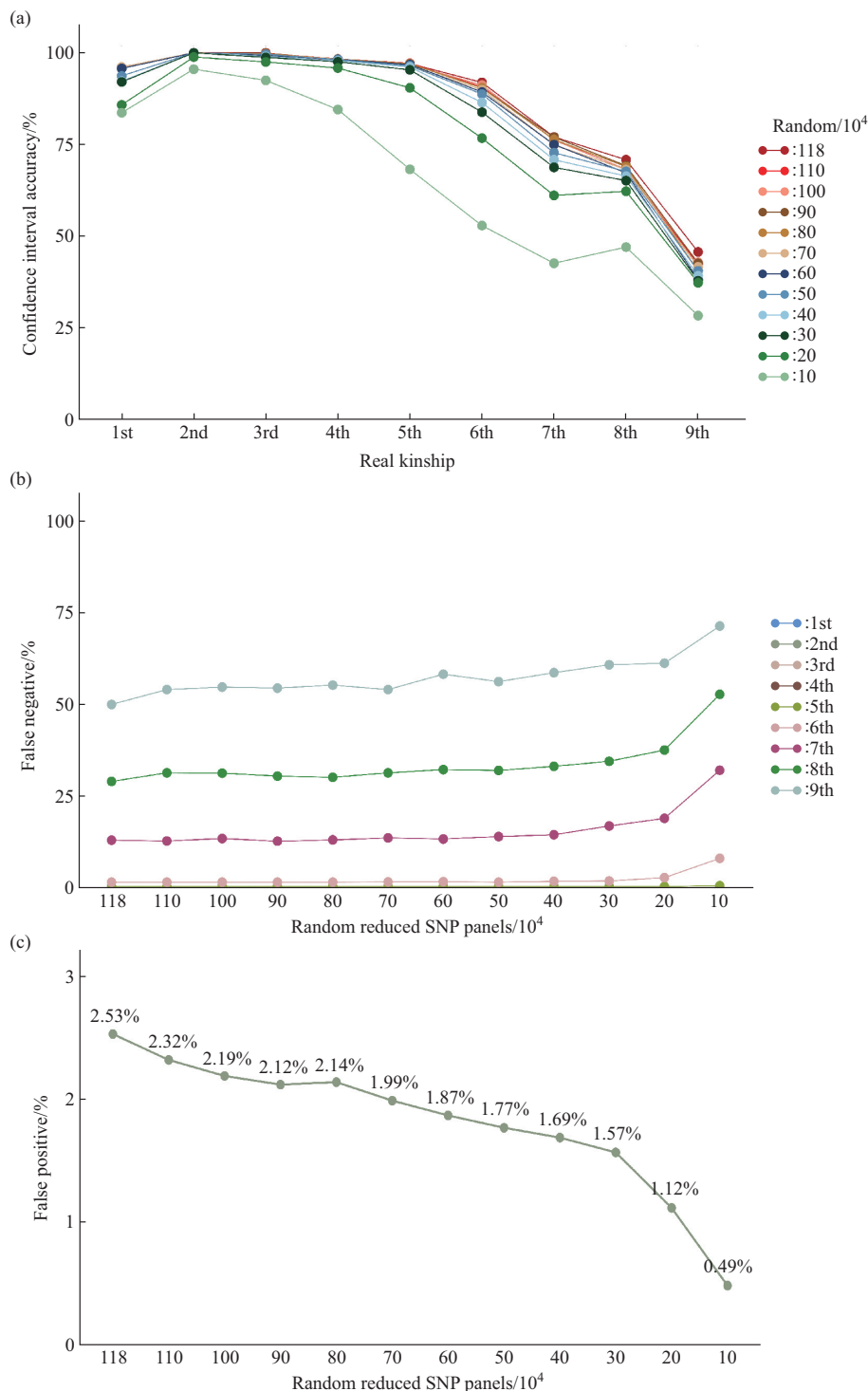


Fig. 5 The prediction accuracy results of randomly reduced loci

2.4 最优IBD推断参数确定

为进一步提升高密度SNP位点集的远亲缘预测能力, 本研究对IBD算法中Germline的错配参数

值(err_hom 、 err_het)按照限定条件(FP 值 $<10\%$)进行筛选。结果显示, 当 $err_het=13$ 、 $err_hom=12$ 时, 亲缘关系预测准确性较高, 将其

标记为P1 (err_het_13_hom_12)。调整参数后的高密度 SNP 位点集, 即P1 (err_het_13_hom_12) 与使用原始错配参数 (err_het_3_hom_2) 的位点集 (P1、P2、P3、P4、P5) 的性能对比情况如下 (图

6)。图6为IBD算法优化后的平均IBD长度对比结果。与原始错配参数的P1、P2、P3、P4、P5结果对比表明, P1 (err_het_13_hom_12) 显著提高了每一等级的平均IBD长度 (秩和检验, $P<0.05$)。

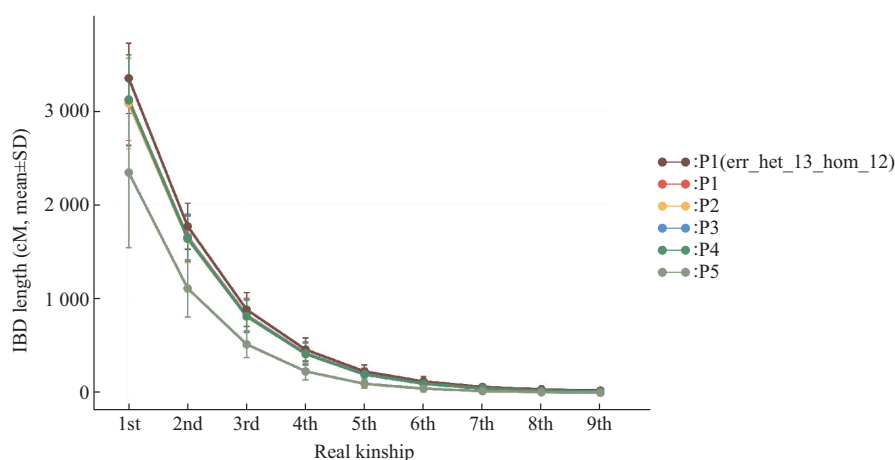


Fig. 6 Comparison results of the average IBD segment length after the optimization of the IBD algorithm (cM)

Data are presented as mean ($\pm$ S.D.). The predicted results of P1, P2, P3, P4, and P5 use the original mismatch parameters (err_het_3_hom_2). err_hom: the maximum number of mismatching homozygous markers for a slice to still be considered part of a match. Default: 1. err_het: the maximum number of mismatching heterozygous markers for a slice to still be considered part of a match. Default: 0.

图7为IBD算法优化后的预测准确性对比结果 (表S4)。表S4和图7的a、b、c分别为AC、CIA、FN、FP的变化情况。在图7a中, 与P1相比, P1 (err_het_13_hom_12) 的1级CIA值提高到100%, 同时提高了4~9级CIA值; 与P2、P3、P4相比, P1 (err_het_13_hom_12) 的5~9级CIA值分别平均提高2.36%、3.92%、8.33%、11.56%、13.06%, 其中显著性提升了8级CIA值 (卡方检验, $P<0.05$); 与P5相比, P1 (err_het_13_hom_12) 的1~9级CIA值均有显著性提升 (卡方检验, $P<0.05$)。在图7b中, 与P1相比, P1 (err_het_13_hom_12) 降低了5~9级FN值; 与P2、P3、P4相比, P1 (err_het_13_hom_12) 的5~9级FN值分别平均降低0.42%、0.53%、2.72%、11.29%、17.57%, 其中显著降低了8~9级FN值 (卡方检验, $P<0.05$); 与P5相比, P1 (err_het_13_hom_12) 的6~9级FN值均有显著性降低 (卡方检验, $P<0.05$)。但是, 在图7c中, P1 (err_het_13_hom_12) 显著增加了FP值 (卡方检验, $P<0.05$)。表明通过调整IBD算法中Germline的错配参数值能显著提升高密度SNP位点集的远亲缘预测能力。

3 讨 论

常用SNP芯片技术对法医DNA的质量有严格要求, 通常要求DNA投入量要达到200 ng以上, 对低质量的法医DNA检测效果较差^[32]。此外, SNP芯片的位点集合具有固定性, 难以借助调整SNP位点的数量来提升亲缘关系预测能力。而从WGS数据中可提取任意SNP位点集用于亲缘关系推断^[28, 33], 具有较强灵活性, 是构建远亲缘关系预测能力更强的SNP位点集合的有效途径。因此, 本研究基于实验室前期研究成果, 并采集了家系人群的WGS数据^[23, 34], 探索能否通过增加SNP位点的数量来提升远亲缘关系的预测能力。本研究合并了3种常用的SNP芯片的位点集, 并进一步筛选出与千人基因组计划中东亚人群位点存在交集的部分, 得到一组高密度SNP位点集。通过将该高密度位点集与3种芯片原始SNP位点集、3种芯片位点的交集及通过随机减少位点数量生成的对照集合进行对比分析, 结果发现, 高密度SNP位点集可以检测到更长的IBD, 从而显著增强了对于远亲缘关系的预测效能。

本研究首先对IBD长度阈值进行筛选, 分析不

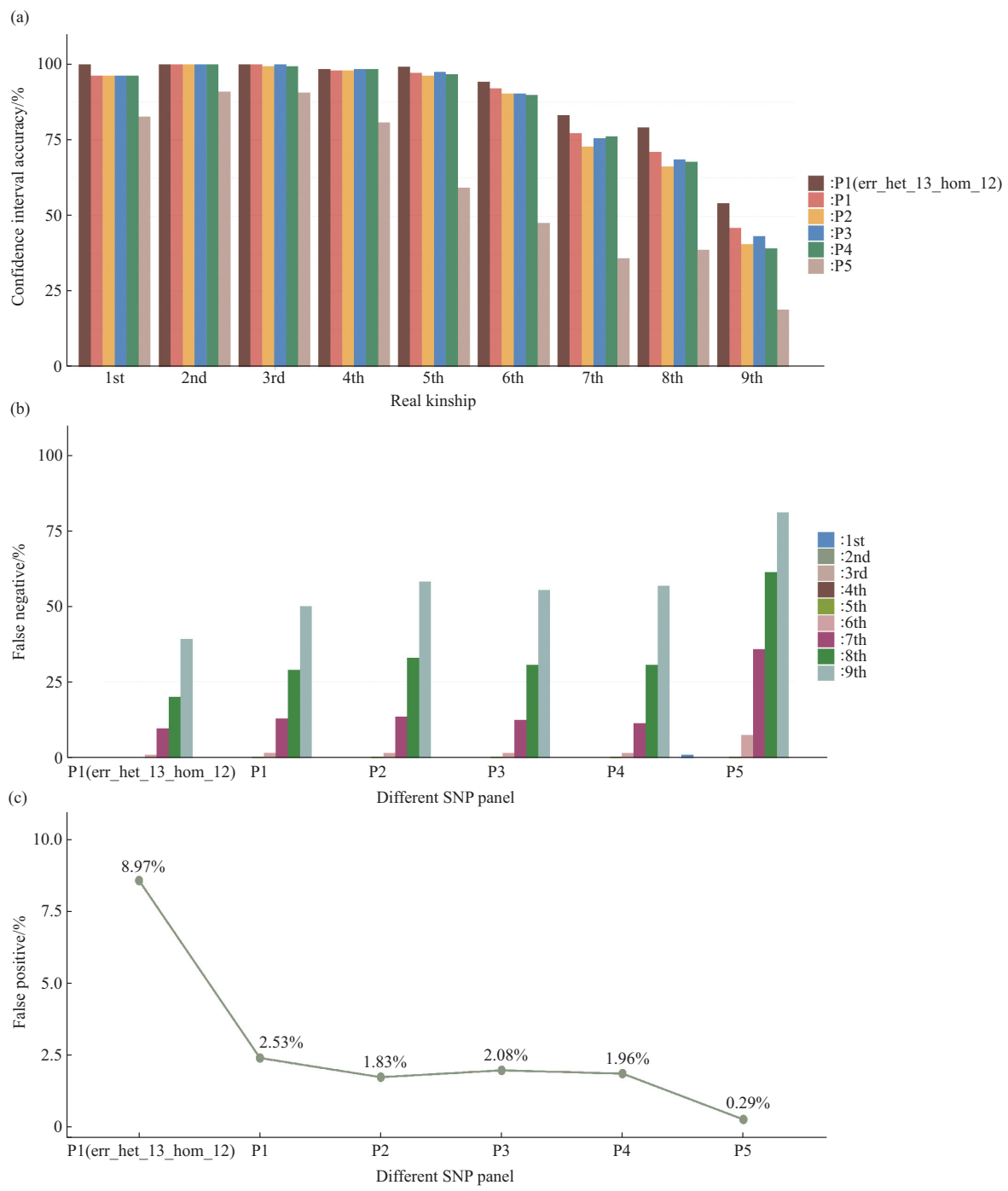


Fig. 7 Comparison results of prediction accuracy after the optimization of the IBD algorithm

The predicted results of P1, P2, P3, P4, and P5 use the original mismatch parameters (err_het_3_hom_2). Confidence interval accuracy (CIA): the probability that two related individuals is predicted to be within the range of 1st of their survey kinship; false negative (FN): the probability of predicting two related individuals as unrelated; false positive (FP): the probability of predicting two unrelated individuals as within 9th degree.

同cM 阈值对亲缘推断的影响。研究结果呈现于图 2, 显示预测准确性随 cM 阈值增加而降低, 鉴于亲缘关系越远, 个体间共 IBD 越短, 远亲缘关系受 cM 阈值筛选的影响程度越大^[20]。其中 1 级 CIA 值呈降低趋势, 是因为存在 1 级被预测为 3 级的关系

对, 而且此类预测错误的关系对数量也随预测等级的提升而增多。刘京等^[23] 利用家系芯片数据进行 cM 阈值分析的过程中, 同样出现了 1 级被预测为 3 级的现象, 并且 IBD 算法在对中国人进行亲缘分析时 2 级的 AC 值较低。究其原因, IBD 算法分析

软件的构建和验证是以欧美人群为基础^[35],一定程度上影响了该软件在分析其他地区人群时的准确性。后续将通过大量中国人群家系数据训练算法,进一步提高IBD算法在中国人群中的预测准确性。

对高密度SNP位点集与3种芯片原始SNP位点集、3种芯片位点的交集以及随机减少位点集的亲缘推断结果展开对比。依据IBD长度结果(表2,图4),发现高密度SNP位点集提升了IBD长度。鉴于位点数量增多,可从样本中获取更多数量及更长长度的IBD,使预测到的关系对也会增加,此从本质上阐明了高密度SNP位点集可以提高远亲缘关系的预测能力。对预测准确性结果(图3,5)进行分析后发现,高密度SNP位点集能够有效提升亲缘关系预测的CIA值,尤其是6~9级远亲缘关系的CIA值;同时,该位点集可降低亲缘关系预测的FN值。另外,由表S3可知,其1~8级的PC值平均可达97%以上。

本研究预测结果与刘京等^[23]CGA、GSA芯片数据合并的预测结果相比,高密度SNP位点集在6、7级CIA值分别提高2.68%、5.27%,FN值分别降低2.7%、14.36%。这表明高密度SNP位点集在远亲缘预测领域展现出显著优势,显著提升了远亲缘关系的预测效能,为相关研究提供了更为有力的支持。Al-Khudhair等^[35]通过WGS数据探索发现,增加SNP位点可以提高远亲缘的预测能力,本研究结果与文献报道相一致。本研究通过调整IBD算法中Germline的错配参数值能显著提高IBD的检出能力,以及显著提升8~9级亲缘预测准确性(图6,7)。但是,SNP位点密度提高以及Germline错配参数的调整,均导致FP值升高。主要的原因可能是存在位点连锁或假IBD^[36]等问题。因此本研究拟从WGS数据层面,不受限于现有SNP芯片位点组合,通过连锁不平衡、最小等位基因频率等条件,随机筛选高密度SNP位点集,从而获得一组远亲缘关系预测的CIA值高,而FN和FP值低的位点组合。另一个产生假阳性的原因可能是有些远亲缘关系对很难调查清楚,这与刘京等^[23]在分析家系亲缘关系时发现的问题相符。

4 结 论

本研究合并3种常用的SNP芯片位点集,获得了一组高密度SNP位点集,可实现1~8级亲缘关系的预测。结果发现,与3种SNP芯片位点集及随机减少位点集等结果相比较,其显著提升了远亲缘关

系的预测能力,通过筛选,确定了最适合该位点集的IBD长度阈值及错配参数。基于此,本实验室计划开发以测序技术为基础的靶向测序方案,提升检测能力和计算精度,同时兼顾与不同SNP芯片的兼容性。此成果有望推动法医SNP系谱推断技术向更高精度、更广适用性发展,为系谱学研究提供更可靠工具,在复杂亲缘关系鉴定等领域发挥更大作用。

附件 见本文网络版(<http://www.pibb.ac.cn>, <http://www.cnki.net>):

PIBB_20250340_Table_S1.xlsx

PIBB_20250340_Table_S2.xlsx

PIBB_20250340_Table_S3.xlsx

PIBB_20250340_Table_S4.xlsx

PIBB_20250340_Table_S5.xlsx

参 考 文 献

- [1] Maguire C N, McCallum L A, Storey C, *et al.* Familial searching: a specialist forensic DNA profiling service utilising the National DNA Database® to identify unknown offenders *via* their relatives —The UK experience. *Forensic Sci Int Genet*, 2014, **8**(1): 1-9
- [2] Kling D, Phillips C, Kennett D, *et al.* Investigative genetic genealogy: current methods, knowledge and practice. *Forensic Sci Int Genet*, 2021, **52**: 102474
- [3] Fitzpatrick C. *Forensic Genealogy*. Fountain Valley: Rice Book Press, 2005
- [4] Kaiser J. We will find you: DNA search used to nab Golden State Killer can home in on about 60% of white Americans. *Science*, 2018. DOI: 10.1126/science.aav7021
- [5] Kennett D. Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes. *Forensic Sci Int*, 2019, **301**: 107-117
- [6] Katsanis S H. Pedigrees and perpetrators: uses of DNA and genealogy in forensic investigations. *Annu Rev Genomics Hum Genet*, 2020, **21**: 535-564
- [7] Dowdeswell T L. Forensic genetic genealogy: a profile of cases solved. *Forensic Sci Int Genet*, 2022, **58**: 102679
- [8] Bieber F R, Brenner C H, Lazer D. Finding criminals through DNA of their relatives. *Science*, 2006, **312**(5778): 1315-1316
- [9] Greytak E M, Moore C, Armentrout S L. Genetic genealogy for cold case and active investigations. *Forensic Sci Int*, 2019, **299**: 103-113
- [10] Greytak E, Wyatt S, Cady J, *et al.* Investigative genetic genealogy for human remains identification. *J Forensic Sci*, 2024, **69**(5): 1531-1545
- [11] Tillmar A, Sjölund P, Lundqvist B, *et al.* Whole-genome sequencing of human remains to enable genealogy DNA database searches - a case report. *Forensic Sci Int Genet*, 2020, **46**: 102233

- [12] Tillmar A, Fagerholm S A, Staaf J, *et al.* Getting the conclusive lead with investigative genetic genealogy - a successful case study of a 16 year old double murder in Sweden. *Forensic Sci Int Genet*, 2021, **53**: 102525
- [13] 刘京, 郭志芳, 魏以梁, 等. 法医 SNP 系谱推断技术破获古汉墓被盗案 1 例. *法医学杂志*, 2021, **37**(4): 600-602
Liu J, Guo Z F, Wei Y L, *et al.* *J Forensic Med*, 2021, **37**(4): 600-602
- [14] 刘京, 马咪, 魏以梁, 等. 法医 SNP 系谱推断技术助破 14 年久冷案. *刑事技术*, 2021, **46**(6): 652-656
Liu J, Ma M, Wei Y L, *et al.* *Forensic Sci Technol*, 2021, **46**(6): 652-656
- [15] 刘京, 何传锦, 张科, 等. 法医 SNP 系谱推断技术助破 23 年冷案 1 例. *法医学杂志*, 2023, **39**(3): 312-314
Liu J, He C J, Zhang K, *et al.* *J Forensic Med*, 2023, **39**(3): 312-314
- [16] Xu Y, Wang N, Gao S, *et al.* Solving the two-decades-old murder case through joint application of ZooMS and ancient DNA approaches. *Int J Leg Med*, 2023, **137**(2): 319-327
- [17] Kling D. On the use of dense sets of SNP markers and their potential in relationship inference. *Forensic Sci Int Genet*, 2019, **39**: 19-31
- [18] Tillmar A, Sturk-Andreaggi K, Daniels-Higginbotham J, *et al.* The FORCE panel: an all-in-one SNP marker set for confirming investigative genetic genealogy leads and for general forensic applications. *Genes*, 2021, **12**(12): 1968
- [19] Thompson E A. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 2013, **194**(2): 301-326
- [20] Seidman D N, Shenoy S A, Kim M, *et al.* Rapid, phase-free detection of long identity-by-descent segments enables effective relationship classification. *Am J Hum Genet*, 2020, **106**(4): 453-466
- [21] Weir B S, Anderson A D, Hepler A B. Author Correction: Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet*, 2022, **23**(2): 134
- [22] Ramstetter M D, Dyer T D, Lehman D M, *et al.* Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, 2017, **207**(1): 75-82
- [23] Liu J, Wei Y L, Yang L, *et al.* Testing of two SNP array-based genealogy algorithms using extended Han Chinese pedigrees and recommendations for improved performances in forensic practice. *Electrophoresis*, 2023, **44**(17/18): 1435-1445
- [24] de Vries J H, Kling D, Vidaki A, *et al.* Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy. *Forensic Sci Int Genet*, 2022, **56**: 102625
- [25] Alaeddini R, Walsh S J, Abbas A. Forensic implications of genetic analyses from degraded DNA—a review. *Forensic Sci Int Genet*, 2010, **4**(3): 148-157
- [26] Yagasaki K, Mabuchi A, Higashino T, *et al.* Practical forensic use of kinship determination using high-density SNP profiling based on a microarray platform, focusing on low-quantity DNA. *Forensic Sci Int Genet*, 2022, **61**: 102752
- [27] Turner S D, Nagraj V P, Scholz M, *et al.* Evaluating the impact of dropout and genotyping error on SNP-based kinship analysis with forensic samples. *Front Genet*, 2022, **13**: 882268
- [28] Loreille O, Tillmar A, Brandhagen M D, *et al.* Improved DNA extraction and illumina sequencing of DNA recovered from aged rootless hair shafts found in relics associated with the romanov family. *Genes*, 2022, **13**(2): 202
- [29] 孙一杰, 逯经纬, 李晶, 等. 基于全基因组测序技术的法医 SNP 系谱推断灵敏度研究. *刑事技术*, 2025. <https://doi.org/10.16467/j.1008-3650.2025.0014>
Sun Y J, Lu J W, Li J, *et al.* *Forensic Sci Technol*, 2025. <https://doi.org/10.16467/j.1008-3650.2025.0014>
- [30] McKenna A, Hanna M, Banks E, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 2010, **20**(9): 1297-1303
- [31] Gusev A, Lowe J K, Stoffel M, *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res*, 2009, **19**(2): 318-326
- [32] Glynn C L. Bridging disciplines to form a new one: the emergence of forensic genetic genealogy. *Genes*, 2022, **13**(8): 1381
- [33] Williams C M, Scelza B A, Slack S D, *et al.* A rapid accurate approach to inferring pedigrees in endogamous populations. *Genetics*, 2025, **230**(4): iyaf094
- [34] Xie Q, Zhao W, Liu W, *et al.* Forensic SNP genealogy inference using whole genome sequencing data of varying depths. *Forensic Sci Int Genet*, 2025, **79**: 103296
- [35] Al-Khudhair A, Qiu S, Wyse M, *et al.* Inference of distant genetic relations in humans using “1000 genomes”. *Genome Biol Evol*, 2015, **7**(2): 481-492
- [36] Li H, Glusman G, Hu H, *et al.* Relationship estimation from whole-genome sequence data. *PLoS Genet*, 2014, **10**(1): e1004144

Research on The Genealogical Inference Efficiency of High-density SNPs*

LI Jing^{1, 3)**}, SUN Yi-Jie^{2, 3)**}, ZHAO Wen-Ting³⁾, TANG Zi-Chen⁴⁾, LIU Jing^{3)***}, LI Cai-Xia^{1, 2, 3)***}

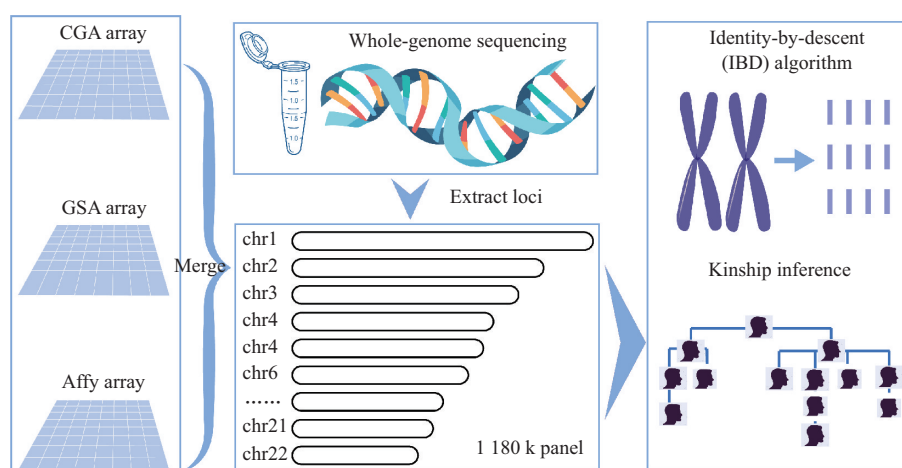
⁽¹⁾College of Forensic Medicine, Hebei Key Laboratory of Forensic Medicine, Collaborative Innovation Center of Forensic Medical Molecular Identification, Hebei Medical University, Shijiazhuang 050017, China;

⁽²⁾Institute of Criminal Investigation, People's Public Security University of China, Beijing 100038, China;

⁽³⁾Key Laboratory of Forensic Genetics, Beijing Engineering Research Center of Crime Scene Evidence Examination, Institute of Forensic science, Beijing 100038, China;

⁽⁴⁾Jiangsu Key Laboratory of Phylogenomics and Comparative Genomics, School of Life Sciences, Jiangsu Normal University, Xuzhou 221000, China)

Graphical abstract



Abstract Objective This study aims to explore the potential of different orders of magnitude single-nucleotide polymorphism (SNP) locus combinations for predicting distant kinship relationships. A high-density SNP locus set was constructed, and a comprehensive assessment of its inference capability was conducted. **Methods** Firstly, we selected three commercial chip panels, CGA (Chinese genotyping array, Illumina), GSA (Global screening array, Illumina), Affy (23MF_V2 high-density SNP array, Affymetrix) and merged them after quality control, forming a high-density SNP locus panel (1 180 k). Secondly, we selected 161 samples and collected their peripheral blood samples by using whole-genome sequencing technology. Within this sample population, the levels of kinship relationships fully covered the range from level 1 to level 9, and the number of kinship pairs at each level was consistently maintained at over 50 pairs. From 161 samples data of whole-genome sequencing, the 1 180 k locus set was extracted, which is referred to as the high-density SNP locus set in the following text. The kinship inference was conducted using the identity-by-descent (IBD) algorithm with the

* This work was supported by grants from the Fundamental Research Fund for Institute of Forensic Science (2024JB026, 2024JB043), The National Natural Science Foundation of China (82171870), and the National Key Research and Development Program of China (2022YFC3341004).

** These authors contributed equally to this work.

*** Corresponding author.

LI Cai-Xia. Tel: 86-10-83752706, E-mail: 90650101@hebmh.edu.cn

LIU Jing. Tel: 86-18211057899, E-mail: biojing@yeah.net

Received: July 18, 2025 Accepted: January 4, 2026

selected optimal parameters. To comprehensively evaluate the performance of the high-density SNP locus set in kinship inference, we compared it with the three commercial chip panels, the intersection of these three chip loci, and the control sets constructed by randomly reducing the number of the high-density SNP locus set. Based on the changes in the IBD lengths, as well as the dynamic trends in prediction accuracy, we conducted a scientific assessment of the kinship inference capability of the high-density SNP locus set. **Results** After screening, a set of 1 184 334 autosomal SNPs was obtained. During the process of screening the optimal IBD length threshold, the result revealed that 0 cM, 1 cM, and 2 cM all demonstrated good applicability. However, to avoid the issue of a large amount of redundant information caused by setting a too low IBD length threshold, this study ultimately selected 2 cM as the optimal threshold. Compared with the average results of three chip panels, the high-density SNP locus set increased the total IBD length and the average IBD length across levels 1–9; the accuracy of the confidence interval for level 8 was 70.97%, which represented a 3.50% improvement; the average confidence interval accuracy for levels 1–8 was 91.39%, representing a 1.00% increase; and the false negative rates at levels 8 and 9 were reduced by 2.42% and 6.76%, respectively. The system efficacy of the high-density SNP locus set for kinship inference of first to eighth degree relationships reached 98.91%. Through random reduction of the high-density SNP locus set results, it is found that increasing the number of SNPs with the panel, the detection efficiency of IBD length showed a significant upward trend. At the same time, the overall trend in the accuracy of kinship relationship prediction as well as the confidence interval accuracy also indicated that both metrics steadily increased with the addition of more loci. **Conclusion** The results show that the high-density SNPs panel significantly enhances the efficacy of distant kinship inference, accurately covering kinship degrees, with the average confidence interval accuracy for first to eighth degree relationships stably above 90%. The study finds that increasing the number of SNPs panel can improve the ability to predict distant kinship.

Key words high-density SNPs panel, forensic SNP genealogy inference, whole genome sequencing, genealogy inference, identity by descent algorithm

DOI: 10.3724/j.pibb.2025.0340

CSTR: 32369.14.pibb.20250340