



基于高密度单核苷酸多态性的共祖远亲缘关系预测算法准确性研究*

刘京^{1,2)} 李晶²⁾ 杨澜²⁾ 管珊珊²⁾ 魏以梁³⁾ 赵雯婷²⁾ 江丽²⁾
赵东^{1)**} 李彩霞^{2)**}

⁽¹⁾ 中国政法大学, 证据科学教育部重点实验室, 北京 100088;

⁽²⁾ 公安部鉴定中心, 法医遗传学公安部重点实验室, 现场物证溯源技术国家工程实验室, 北京 100038;

⁽³⁾ 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 221116)

摘要 目的 研究构建基于共祖 (identity-by-descent, IBD) 片段算法预测远亲缘关系分析流程并评估预测准确性。**方法** 采用高密度单核苷酸多态性 (single nucleotide polymorphism, SNP) 芯片对 253 份家系样本进行检测, 研究基于 IBD 片段算法的分析流程进行两两个体间亲缘关系预测, 评估预测准确性。随机减少 SNP 位点, 评估位点数对算法预测准确性的影响。**结果** IBD 片段算法预测 1~7 级亲缘关系平均置信区间准确率为 94.72%, 预测可信度为 99.77%, 6 级及以上亲缘关系预测时出现假阴性。随着 SNP 数量减少, 预测准确性会出现一定程度的下降。**结论** IBD 片段算法可用于 7 级以内亲缘关系的预测, 该算法在群体遗传学、法医遗传学等领域有重要应用价值。

关键词 共祖片段算法, 法医系谱推断, 远亲缘关系预测

中图分类号 R89, Q812

DOI: 10.16476/j.pibb.2022.0538

短串联重复序列 (short tandem repeat, STR) 是司法鉴定领域进行亲缘关系鉴定的主要遗传标记, 但其只能对父子^[1]、同胞^[2]、祖孙^[3]等 2 级以内的近亲缘进行关系鉴定。单核苷酸多态性 (single nucleotide polymorphism, SNP) 位点具有分布广泛、突变率低等特点, 随着全基因组测序 (whole genome sequencing, WGS)、高密度 SNP 基因芯片等检测技术的发展成熟, 利用高密度 SNP 数据预测远至 7~9 级亲缘关系成为近年来法医遗传学领域研究热点^[4-10], 该技术即为法医 SNP 系谱推断技术。2018 年美国警方首次使用法医 SNP 系谱推断技术搜索到“金州杀手”的远亲^[11], 然后构建系谱树, 进而找到凶手成功破获了 42 年前的冷案, 该技术被《科学》杂志 (Science) 评为当年十大科学突破之一。此后, 美国警方利用该技术为数百起冷案积案提供关键线索。一项基于美国白人的研究表明^[4], 建立约占人口 2% 的 SNP 数据库, 即可为约 99% 的人口找到至少一名 3 代表/堂亲 (即 7 级亲缘)。研究和实践表明^[12-14], 法医 SNP 系谱推断与传统 STR 技术相结合已成为法医 DNA 服

务冷案积案侦查的新模式。

目前基于高密度 SNP 数据进行远亲缘关系推断的方法众多, 包括: 似然比算法、共享等位基因比例 (identical by state, IBS) 算法、共祖 (identity-by-descent, IBD) 片段算法等^[15]。似然比算法需提前给出一对个体之间具体关系类别的备择假设和两者无亲缘的原假设, 再根据亲缘关系的两个互斥假设, 观察个体间的遗传标记数据的条件概率, 通过比较得到两假设条件概率的似然比。IBS 算法^[16-17]通过评估样本间每个 SNP 等位基因频率, 计算基因组中共享等位基因比例程度确定亲缘关系等级。以上两种算法适合预测 4 级以内亲缘关系, 5 级以上的亲缘关系预测准确性显著降低^[16]。IBD 片段算法通过检测 IBD, 即来自一个共

* 国家自然科学基金 (82171870), 公安部技术研究计划 (2021JSZ15) 和中央级公益性科研院所基本科研业务费 (2021JB004) 资助项目。

** 通讯联系人。

李彩霞 Tel: 010-83752706, E-mail: licaixia@tsinghua.org.cn

赵东 Tel: 15010055761, E-mail: dzfslm@163.com

收稿日期: 2022-11-21, 接受日期: 2023-04-11

同祖先的相同DNA片段长度和数量,判断亲缘关系远近。由于减数分裂时,父母双方的DNA会发生断裂和重组,亲缘关系越远意味着传递代数(重组)越多,个体间共享的IBD片段就越短,IBD长度使用厘摩(cM)衡量。该算法适于预测7级左右的亲缘关系,某些情况下可预测高达12级的亲缘关系^[18]。

IBD片段算法是目前法医系谱推断最常用算法。但在中国法医遗传学应用实践中,存在如下问题:a.国外研究大多基于模拟亲缘关系对和欧美人群亲缘关系对^[18],欠缺适合中国人群真实亲缘关系的参数优化、准确性评估等系统性研究;b.需对高密度SNP数据进行格式转换、同源染色体分型等处理,分析流程繁琐,普通法医工作者难以完成。本项目组在国内首次研究构建了适合法医应用的IBD亲缘关系级预测的自动分析算法流程,实现了大量样本两两个体之间亲缘关系的批量计算,并为多起命案积案侦破提供了关键线索^[12-13]。本文详述了该算法流程的构建和优化研究,并基于中国5个汉族大家系样本的真实亲缘关系对进行了IBD片段算法预测准确性评估。本文研究成果将为中国法医SNP系谱推断技术的研究和应用提供数据支撑和软件支持。

1 材料与方法

1.1 样本来源

采集本研究团队5名成员的汉族家系253人份样本,总共5560对亲缘关系(图1),包括一级

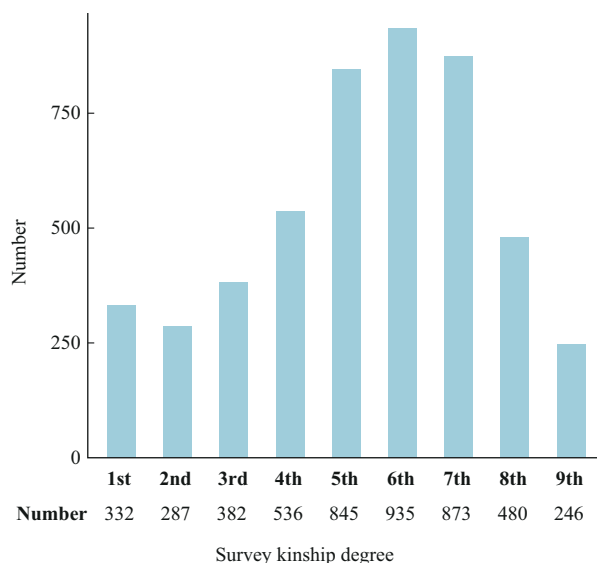


Fig. 1 The distribution of survey kinship degree

(1st)、二级(2nd)、三级(3rd)、四级(4th)、五级(5th)、六级(6th)、七级(7th)、八级(8th)、九级(9th)、大于九级(10~14级, >9th)亲缘关系,以及26318对无亲缘关系。所有样本在采集前均签署知情同意书,本研究通过了公安部鉴定中心伦理委员会审查(编号:2021-006)。

1.2 DNA提取与检测

使用MagAttract M48 DNA Manual试剂盒(Qiagen公司,德国)提取DNA,使用NanoDrop 2000c分光光度计(Thermo Scientific公司,美国)进行DNA定量。使用WeGene V2基因芯片(安澜智能公司,中国)进行SNP检测(DNA模板量均大于500 ng,芯片位点检出率均大于98.5%),获得约70万SNP位点分型数据。

1.3 数据分析

本研究形成的IBD片段算法集成到项目组前期开发的DNA系谱推断系统DGA v1.0^[19]进行亲缘关系预测。

1.4 亲缘关系推断算法参数优化

1.4.1 预测准确性评估指标

为了评估IBD片段算法在中国真实家系亲缘关系对预测准确性,本文使用绝对准确率(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)、预测可信度(prediction credibility, PC)作为评估指标。AC是指某亲缘等级的调查亲缘关系对应的所有关系对中,预测结果同样是此等级的关系对所占的比例;CIA是指某亲缘等级的调查亲缘关系对应的所有关系对中,预测结果是此等级或此等级 ± 1 级的关系对所占的比例^[16];FN是指某亲缘等级的调查亲缘关系对应的所有关系对中,预测结果是“无关”的关系对所占的比例;FP是调查亲缘关系为“无关”(大于14级)的所有关系对中,预测结果是“有关”(1~9级)的关系对所占的比例;PC是指某亲缘等级的预测亲缘关系对应的所有关系对中,调查亲缘关系为“有关”的关系对所占的比例。

1.4.2 IBD片段长度阈值优化

为了评估不同匹配片段最低检出长度阈值对预测准确性的影响,设置了0、3、6、9、12、15、20 cM 7个不同的IBD片段长度阈值,评估不同最低检出IBD片段长度阈值的预测准确性。

1.4.3 支持向量机 (SVM) 优化

为了提高亲缘关系预测等级的准确性,本研究使用支持向量机 (support vector machine, SVM) 方法对 IBD 片段算法中基于 IBD 片段数量和长度预测亲缘关系等级进行优化,将预测有亲缘关系 IBD 片段总长度和平均长度的最小值设为阈值。本研究根据阈值设置 3 个集合 (图 2): 集合 1, 预测为有亲缘关系的关系对; 集合 2, 有 IBD 片段结果的关系对中 IBD 片段总长度和平均长度均小于该阈值的关系对; 集合 3, 有 IBD 片段结果的关系对中 IBD 片段总长度和平均长度均大于该阈值但被预测为无关的关系对, 将集合 1 和集合 2 合并作为训练集, 将集合 3 作为测试集, 通过训练集对测试集重新进行分类, 以降低较远 (6 级及以上) 亲缘关系预测的 FN。

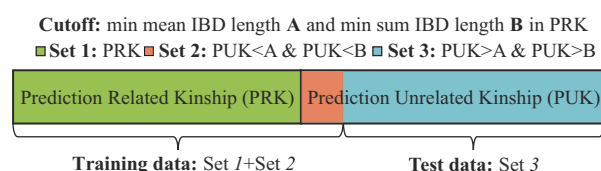


Fig. 2 The SVM training and test data set

1.5 不同数量 SNP 预测评估

由于法医物证受时间、环境等影响,经常会发生 DNA 降解,检出的 SNP 位点会随之减少。为了模拟降解 DNA 对该系谱推断算法预测效能影响,本文对位点进行随机的梯度下降筛选,将筛选的位点组合进行两次亲缘关系预测的平均结果与原始数据结果进行比较,评估不同密度 SNP 位点组合对预测准确性的影响。

2 结 果

2.1 基于 IBD 片段算法的亲缘关系分析流程

本文研究的 IBD 片段算法包括如下分析流程: 过滤 SNP 数据中的冗余信息,筛选位点,对筛选后数据进行格式转换,同源染色体分离,查找和合并各染色体上 IBD 匹配片段,基于 IBD 片段长度、数量等值预测个体间亲缘关系等级,在算法研制过程中进行 IBD 片段长度阈值、SVM 亲缘等级预测等优化 (图 3)。其中同源染色体分离选取了千人基因组中国人人群作为参考数据集,使用隐马尔可夫 (HMM) 算法将待分析数据父源和母源染色体分离; IBD 片段长度计算时参考了 HapMap 计划网站

中 SNP 物理距离 (bp) 与遗传距离 (cM) 之间的关系进行厘摩值转换。基于以上分析流程,本研究使用 Python 编写了基于 IBD 片段算法进行亲缘关系分析的 pipeline,实现了数据预处理自动化、两两个体亲缘关系计算批量化等功能。

Algorithm: IBD algorithm analysis process

Data: plink data U: bed, bi, fam

Result: predict kinship degree

```
1 We filter plink data U and convert U to vcf format V;  
2 We phase data V and get phased data VP; then we convert  
   phased VP to plink data U1;  
3 We impute plink data U1 with genetic diactanc and get plink  
   data U2; then, we set IBD length threshold L;  
4 for i ∈ L do  
5   Minimum IBD length: i;  
6   Infer pairwise IBD segment from plink data U2;  
7   Estimate kinship degree by IBD segment;  
8 End  
9 Finally, select optimal minimum IBD length i and use SVM to optimize  
   the prediction kinship degree.
```

Fig. 3 IBD algorithm analysis process

2.2 亲缘关系推断参数及 SVM 优化结果

2.2.1 亲缘关系推断算法结果

使用上述 IBD 片段算法对 253 份高密度 SNP 数据进行亲缘关系计算,将所有个体间预测的 1~9 级亲缘关系等级与实际调查的亲缘关系进行比较并计算准确性评估指标 (表 1)。表 1 可以看出,前 5 级有较高的准确率,平均 CIA 为 99.14%, FN 为 0。随着亲缘关系等级的增加,准确率也随之降低,6 级开始出现假阴性,8 级及以上假阴性明显增加。1~7 级亲缘的预测可信度较高,平均值为 99.75%。

2.2.2 IBD 片段长度阈值研究结果

使用软件计算两两个体 IBD 片段长度时,需设置最低检出 IBD 片段长度阈值。为评估该参数对预测准确性影响,本文设置了 0 (即无最低检出 IBD 片段长度限制)、3 (默认参数)、6、9、12、15、20 cM 七个不同阈值长度阈值,计算在 253 份样本中预测准确性变化情况。图 4a 展示了不同 IBD 片段长度阈值下 AC 变化情况;图 4b 展示了不同 IBD 片段长度阈值下 CIA 变化情况;图 4c 展示了不同 IBD 片段长度阈值下 PC 变化情况;图 4d 展示了不同 IBD 片段长度阈值下 FN 变化情况;图 4e 展示了不同 IBD 片段长度阈值下 FN 平均值和 FP 变化情况。对比显示 0~9 cM IBD 片段阈值参数对预测准确性影响不大,当 IBD 片段阈值大于 9 cM, CIA、FP 均有一定程度降低,而 PC、FN 有一定程度增加。

Table 1 Accuracy statistics of IBD algorithm in predicting kinship degree of 253 samples

Survey	Predict Kinship										AC/%	CIA/%	FN/%	FP/%
kinship	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	UN				
1st	331		1								99.70	99.70	0.00	
2nd	2	162	122	1							56.45	99.65	0.00	
3rd			346	33	3						90.58	99.21	0.00	
4th			12	433	91						80.78	100.00	0.00	
5th				58	539	224	20	4			63.79	97.16	0.00	
6th					96	433	294	76	16	20	46.31	88.02	2.14	
7th					6	100	302	228	114	123	34.59	72.16	14.09	
8th						12	58	117	97	196	24.38	56.67	40.83	
9th							3	28	48	167	19.51	30.89	67.89	
>9th							3	10	57	574			89.13	
UN						1	10	74	372	25 861				1.74
PC/%	100	100	100	100	100	99.87	98.41	86.25	44.09					

AC: the probability of correctly predicting two related individuals as the degree of their survey kinship; CIA: the probability that two related individuals is predicted to be within the range of 1st of their survey kinship, and above 7th degree or unknown relationship is not taken into consideration; FN: the probability of predicting two related individuals as unrelated; FP: the probability of predicting two unrelated individuals as within 7th degree.

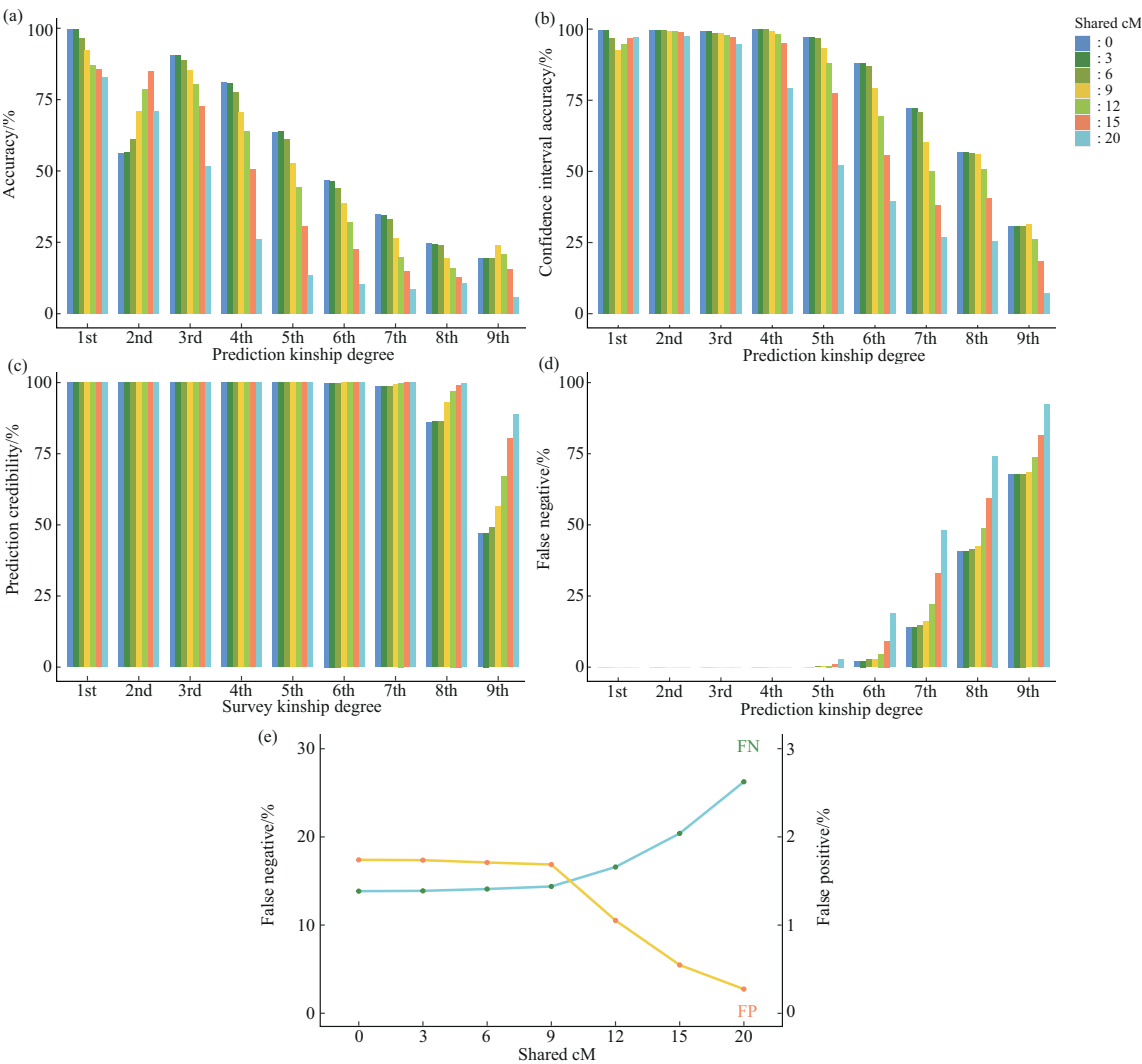


Fig. 4 Accuracy evaluation of different IBD fragments threshold (a-e) corresponding to AC, CIA, PC, FN, FP in different IBD fragments threshold.

2.2.3 SVM优化模型结果

表1结果显示6级及以上亲缘关系预测时会出现假阴性。为提高算法预测能力,降低FN,本研究使用SVM方法对基于IBD片段数量和长度预测亲缘关系等级的过程进行优化,将所有个体间预测的1~9级亲缘关系等级与实际调查的亲缘关系进行比较并计算准确性评估指标(表2)。优化后前5级

平均CIA为99.16%,6级FN由优化前的2.14%降为0.43%,7级FN由优化前的14.09%降为7.10%。为进一步评估SVM优化后预测性能,绝对准确率等评估指标与2.2.1常规流程进行横向对比(图5),结果显示SVM优化后提高了远亲缘(6~9级)关系等级CIA,降低了其FN。

Table 2 Accuracy statistics of SVM IBD algorithm in predicting kinship degree of 253 samples

Survey kinship	Predict Kinship										AC/%	CIA/%	FN/%	FP/%
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	UN				
1st	331		1								99.70	99.70	0	
2nd	2	161	123	1							56.10	99.65	0	
3rd			346	33	3						90.58	99.21	0	
4th			12	434	90						80.97	100	0	
5th				61	537	224	19	4			63.55	97.28	0	
6th					96	436	292	82	25	4	46.63	88.13	0.43	
7th					6	101	303	240	161	62	34.71	73.77	7.10	
8th						12	58	125	152	133	26.04	69.79	27.71	
9th							3	32	77	134	31.30	44.31	54.47	
>9th							3	14	117	510			79.19	
UN						1	10	254	931	25 122				4.54
PC/%	100	100	100	100	100	99.87	98.55	66.18	36.36					

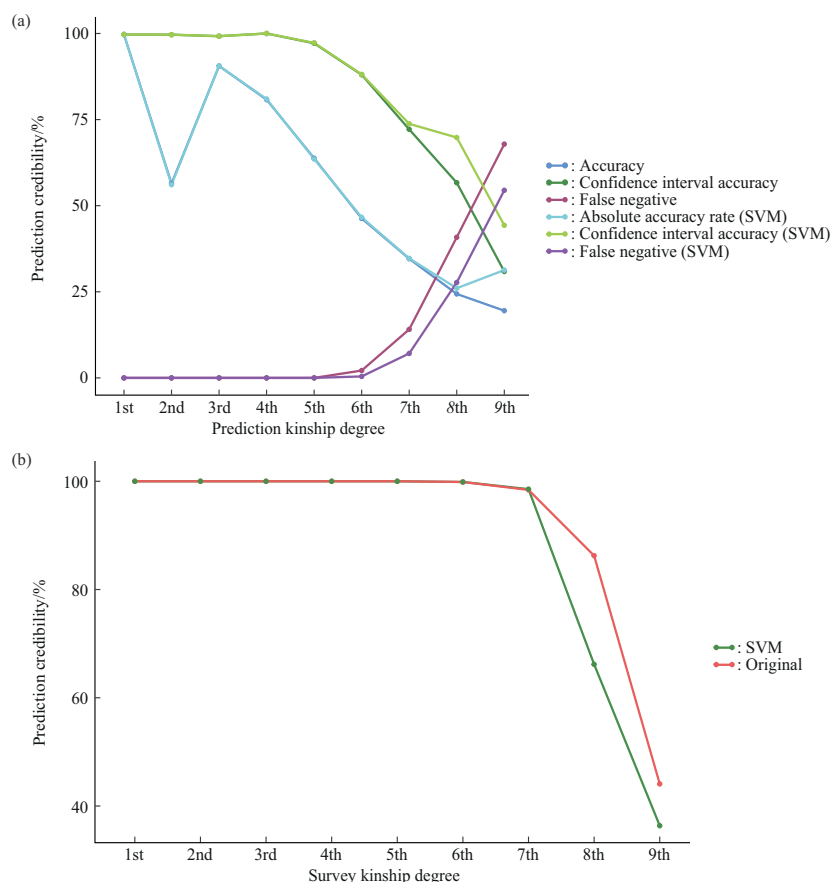


Fig. 5 Accuracy evaluation before and after SVM optimization

(a) AC/CIA/FN rate before and after SVM. (b) PC rate before and after SVM.

2.3 真实亲缘关系IBD片段长度分布

通过研究构建的算法流程, 本文分析253份样本所有真实亲缘关系对的IBD片段长度分布情况。图6显示了253份样本之间真实亲缘关系等级的

IBD片段长度分布图, 可以看出1~4级关系能明显分离开来, 5级以上亲缘关系IBD片段长度分布出现重叠, 亲缘关系越远长度分布的重叠越多。

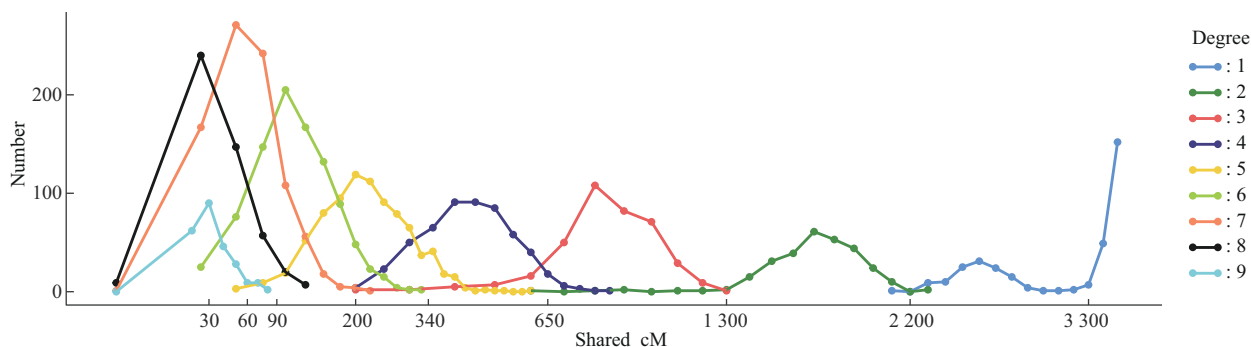


Fig. 6 IBD fragment length statistics for each actual kinship degree of 253 samples

2.4 不同SNP数量预测准确性

生物检材受时间和环境等因素影响, DNA会发生降解, SNP位点检出数也会随之降低。因此本文通过随机筛选不同数量的位点组合, 模拟低质量样本的预测结果。从253份样本数据的所有SNP位点中, 随机筛选65万~10万、每组递减5万位点共

12组, 每组位点个数随机取2次, 使用优化后IBD片段算法预测亲缘关系, 计算AC、CIA等评估指标的2次平均值。不同数量位点预测准确性趋势如图7所示。图7a展示了不同SNP数量AC变化情况; 图7b展示了不同SNP数量CIA变化情况; 图7c展示了不同SNP数量PC变化情况; 图7d展示了

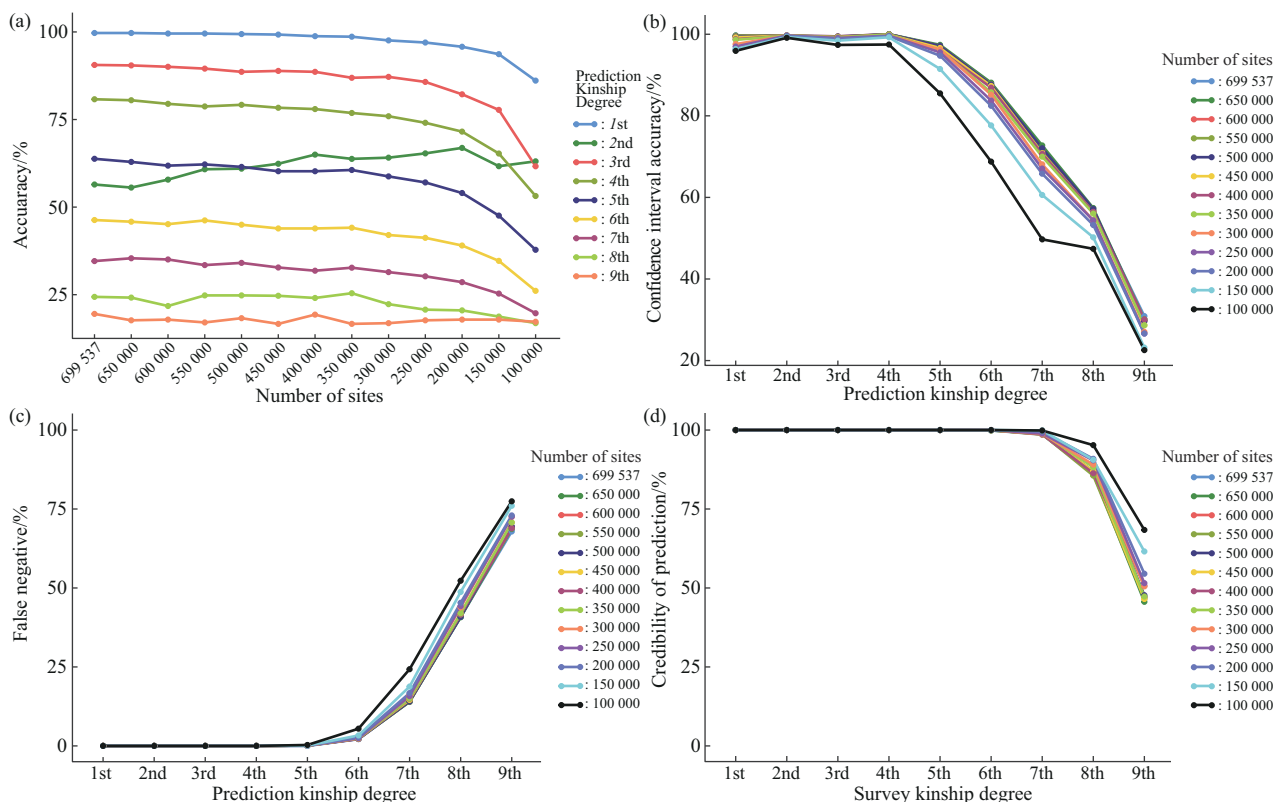


Fig. 7 Accuracy evaluation of different SNP number

(a-d) corresponding to AC, CIA, FN, PC in different SNP number.

不同 SNP 数量 FN 变化情况。结果显示, 随着 SNP 数量的降低, 各预测准确性评估指标会有一定程度的下降 (FN 和 FP 是随着 SNP 数量的降低而略升高), SNP 位点数下降对于超过 5 级的亲缘关系预测能力影响更明显。

3 讨 论

在群体遗传学和法医遗传学应用研究中, 使用高密度 SNP 数据进行亲缘关系预测受到越来越多的关注。项目组先前构建了基于 IBS 算法的分析流程, 并评估了其在中国人真实亲缘中预测准确性^[16], 结果显示, 该算法在 4 级以内亲缘关系有较高预测准确性, 5 级及更远亲缘关系预测准确性下降明显。IBD 片段算法可以准确预测 1~7 级亲缘关系, 在法医 SNP 系谱推断中有广泛应用。但国内缺乏 IBD 片段算法分析流程、针对中国人群的算法优化以及真实亲缘关系预测准确性等系统研究。本文搭建了基于 IBD 片段算法预测亲缘关系的全流程, 实现了输入数据格式预处理、批量化计算两两个体间 IBD 片段长度和亲缘关系等级等功能, 并通过软件进行了分析流程的集成自动化。使用该分析流程对 253 份样本两两间 1~9 级亲缘关系进行预测 (表 1), 结果显示 IBD 片段算法可以准确预测 1~7 级亲缘关系, 平均 CIA 为 94.49%。

为进一步提升 IBD 片段算法预测准确性, 本文进行了最低检出 IBD 片段阈值和 SVM 等优化。不同最低检出 IBD 片段长度阈值结果 (图 4) 显示, 0~9 cM 的 IBD 片段阈值参数对预测准确性影响不大, 大于 9 cM 时预测准确性会有一定程度的降低, 分析其原因可能为: 本流程所使用预测亲缘等级算法^[18]根据两个个体间共享 IBD 片段的数量、长度和位置, 使用原假设 (两个体不相关) 与备择假设 (两个体共享有共同祖先) 进行概率比较。原假设中需要考虑群体中所有共享片段长度的均值, 即群体中随机无关个体共享 IBD 片段长度, 故较短 IBD 片段在该算法中会被当作群体背景噪音。Kling 等^[5]研究表明, 最小 IBD 片段阈值最好选取在 3~8 cM 之间; De Vries 等^[20]研究表明, 设置 1~7 cM 最小 IBD 片段阈值, 在 1 000 对 1~5 级模拟亲缘关系中预测准确率基本无变化, 以上研究结论与本文基本一致。本文还观察到亲缘关系越远, 受不同 IBD 片段长度阈值影响越大 (图 4a), 分析其原因可能为: 由于基因重组的随机性, 亲缘关系越远, 两两个体间共祖片段越短, 故受最低检出阈值影响

就越大。结合本文研究结果, 为在实战中尽可能找到多的亲缘关系, 在后续分析中采用的阈值为 3 cM。使用 SVM 优化 IBD 片段算法后 (图 5), 1~5 级亲缘关系预测的 AC、CIA 等评估指标均无显著变化 (ANOVA 方差分析, $P=0.98$), 6 级 FN 从 2.1% 降低至 0.4%, 7 级 FN 从 14.1% 降低至 7.1%, 7~9 级 CIA 也有一定升高, 1~7 级预测亲缘的 PC 经 SVM 优化后依然保持较高准确率, 平均 PC 为 99.77%。综上所述, 原 IBD 片段算法经 SVM 优化后在远亲缘关系的预测能力方面提升显著。本文对优化后的 IBD 片段算法预测结果进一步探究发现, 7 级的 FN 为 7.1%, 8 级、9 级 FN 急剧上升至 27.71%、54.47%。Greytak 等^[6]研究表明, 由于基因重组的随机性, 大约有 10% 的 3 代表亲 (3rd cousin, 即 7 级) 和 50% 的 4 代表亲 (4th cousin, 即 9 级) 没有可检测到的 IBD 片段长度, 与本文观察到的 FN 结果基本一致。原因可能在于全基因组 SNP 芯片检测的位点数量有限, 在全基因组层面位点之间分布距离较大, 导致部分远亲缘关系对的 IBD 片段检测不到; Al-Khudhair 等^[21]使用 WGS 数据的探索研究表明, 若 SNP 位点分布密度增加, 有可能提高 8 级以及更远亲缘的预测能力。5 级及以下 FN 为 0%, 6 级以上出现假阴性, 7 级的 FN 为 7.1%。故系谱推断实践应用时: 5 级及以下家系只需检测一个样本; 6、7 级家系至少检测两个样本, 这样目标样本与至少一个样本确证可检出亲缘关系的概率可达到 99.99% 和 99.50%。在为侦查提供线索过程中我们发现, 更加关注的 CIA 对于 1~7 级亲缘均超 70%, 但是 2 级亲缘关系预测的 AC 却较低, 大量 2 级亲缘被预测到 3 级。分析其原因可能是, 本流程所使用开源的预测亲缘等级算法构建和验证均基于欧美人群^[18], 在其他人群中的 AC 会有所降低。Williams 等^[22]研究发现, 该算法在非洲辛巴族家系人群中, 2 级亲缘 AC 仅为 67%; Ramstetter 等^[23]研究也发现, 在墨西哥家系人群中, 大量 2 级亲缘关系被预测为 3 级亲缘关系。在不同 IBD 片段阈值 AC 结果中, 2 级亲缘 AC 值出现了与其他等级关系对不同的“先增后降”趋势, 其原因也有可能与人群特异性有关。后续, 需要增加中国人群 2 级和 3 级关系对数量, 继续优化算法模型, 提升中国人群中 2 级亲缘预测的 AC。

基于 253 份样本真实亲缘关系的 IBD 片段长度分布 (图 6) 显示, 1~4 级关系 IBD 长度分布区分显著, 5 级以上亲缘关系 IBD 片段分布会出现重

叠, 亲缘关系越远分布重叠越多。分析其原因为: 亲代向子代传递遗传物质时会发生基因片段的断裂与重组, 亲缘关系越远, IBD 片段长度会缩短且具有一定随机性。通过将美国 Ancestry 公司基于 24 362 份欧美样本模拟亲缘关系厘摩长度分布与本文中国汉族人群真实亲缘关系厘摩长度分布比较发现, 1~7 级厘摩分布范围的趋势大致相仿, 但中国汉族人群每一级亲缘关系对 IBD 片段分布范围的上下限数值与之有所不同。表明不同人群的祖先群体规模、婚配模式等的差异, 会影响不同级别亲缘关系对的 IBD 片段长度分布范围。下一步, 我们将通过模拟家系和真实家系结合的方式, 增加亲缘关系对数量, 进一步验证和优化本研究针对汉族人群研究获得的 IBD 片段分布情况和预测算法, 并研究该体系在蒙古族等其他人群中的适应性。

本研究还进一步通过随机筛选位点数, 模拟低质量样本的系谱推断预测结果。结果显示, 预测准确性随位点数量的减少而降低, 当位点数少于 20 万位点时准确性下降较明显, 但准确性依然处于较高水平, 而且位点数量对 1~3 级近亲缘关系影响更小。但是, 我们模拟随机位点数量减少时没有考虑低检出率 SNP 芯片数据的分型错误率问题。De Vries 等^[20]研究表明, 当 SNP 芯片的检出率降低时, SNP 位点分型的准确率会下降, 而 SNP 分型错误, 会导致 IBD 片段识别提前结束, 造成 IBD 片段丢失, 并最终降低 IBD 片段算法预测准确率, 后续将增加该指标进行模拟数据测试, 并使用真实的低质量检材进行系统的测试。此外, 高深度 WGS 技术可生成同一位置大量短读序列片段 (reads), 从而确保检出 SNP 位点分型的准确性, 针对微量 DNA 有可能获得比 SNP 芯片更加准确的分型结果。

4 结 论

本研究构建了基于高密度 SNP 数据的 IBD 片段算法分析流程并进行了优化, 基于 253 份中国人群样本的真实亲缘关系评估了算法预测准确性。研究结果表明, 该算法可实现 1~7 级亲缘关系的预测。该方法可辅助物证鉴定工作, 为冷案积案等疑难案件侦破提供重要科技支撑。

参 考 文 献

- [1] 中华人民共和国司法部司法鉴定管理局. GB/T 37223-2018, 亲权鉴定技术规范. 北京: 中国标准出版社, 2018: 3-5
- [2] 中华人民共和国司法部司法鉴定管理局. SF/Z JD0105002-2014, 生物学全同胞关系鉴定实施规范. 北京: 中国标准出版社, 2014: 2
- [3] 中华人民共和国司法部司法鉴定管理局. SF/Z JD0105005-2015, 生物学祖孙关系鉴定规范. 北京: 中国标准出版社, 2015: 2-4
- [4] Ministry of Justice of the People's Republic of China. SF/Z JD0105005-2015, Biological grandchild relationship identification norms. Beijing: Standards Press of China, 2015: 2-4
- [5] Erlich Y, Shor T, Pe'er I, *et al.* Identity inference of genomic data using long-range familial searches. *Science*, 2018, **362**(6415): 690-694
- [6] 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
- [7] Greytak E M, Moore C, Armentrout S L. Genetic genealogy for cold case and active investigations. *Forensic Sci Int*, 2019, **299**: 103-113
- [8] 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
- [9] 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
- [10] Dowdeswell T L. Forensic genetic genealogy: a profile of cases solved. *Forensic Sci Int Genet*, 2022, **58**: 102679
- [11] Davawala A, Stock A, Spiden M, *et al.* Forensic genetic genealogy using microarrays for the identification of human remains: the need for good quality samples - a pilot study. *Forensic Sci Int*, 2022, **334**: 111242
- [12] Phillips C. The Golden State Killer investigation and the nascent field of forensic genealogy. *Forensic Sci Int Genet*, 2018, **36**: 186-188
- [13] 刘京, 季安全, 李彩霞, 等. 法医系谱分析研究进展. *刑事技术*, 2019, **44**(3): 189-194
- [14] Liu J, Ji A Q, Li C X, *et al.* *Forensic Sci Technol*, 2019, **44**(3): 189-194
- [15] Kling D, Tillmar A. Forensic genealogy-a comparison of methods to infer distant relationships based on dense SNP data. *Forensic Sci Int Genet*, 2019, **42**: 113-124
- [16] 管珊珊, 张文杰, 魏以梁, 等. 基于 IBS 算法预测亲缘关系准确性研究. *生物化学与生物物理进展*, 2022, **49**(3): 591-599
- [17] Guan S S, Zhang W J, Wei Y L, *et al.* *Prog Biochem Biophys*, 2022, **49**(3): 591-599

- [15] 王鑫,李悦,文豪,等.亲缘关系预测系统.生命科学研究,2020,**24**(3): 229-233+258
Wang X, Li Y, Wen H, *et al.* Life Sci Res, 2020, **24**(3): 229-233+258
- [16] Huff C D, Witherspoon D J, Simonson T S, *et al.* Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Res, 2011, **21**(5): 768-774
- [17] 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
- [18] 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.* Forensic SNP genealogy inference technology solved one case of theft of ancient Han tomb. J Forensic Med, 2021, **37**(4): 600-602
- [19] 张文杰,李彩霞,于迪,等.对未知关系个体进行系谱推断的自动分析系统.生命科学研究,2022,**26**(6): 1-11
Zhang W J, Li C X, Yu D, *et al.* Life Sci Res, 2022, **26**(6): 1-11
- [20] 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
- [21] 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
- [22] Williams C M, Scelza B, Gignoux C R, *et al.* A rapid, accurate approach to inferring pedigrees in endogamous populations. bioRxiv, 2020. doi:10.1101/2020.02.25.965376
- [23] 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

Accuracy Research on The Distant Kinship Relationship Prediction by IBD Algorithm*

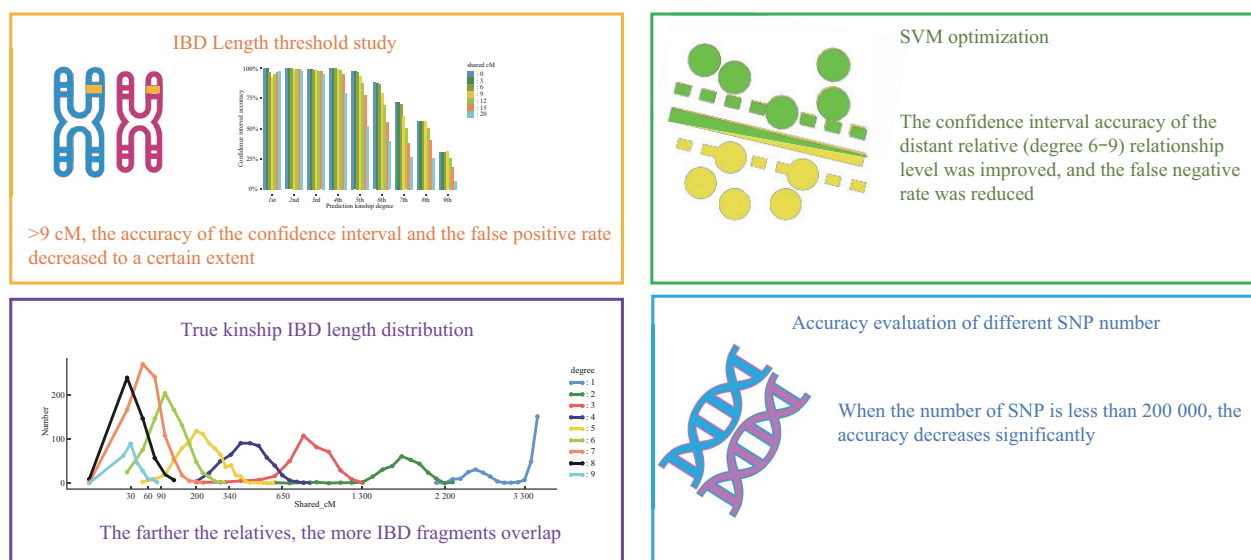
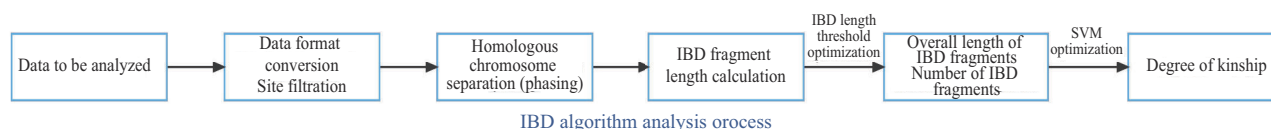
LIU Jing^{1,2)}, LI Jing²⁾, YANG Lan²⁾, GUAN Shan-Shan²⁾, WEI Yi-Liang³⁾, ZHAO Wen-Ting²⁾,
JIANG Li²⁾, ZHAO Dong^{1)**}, LI Cai-Xia^{2)**}

⁽¹⁾Key Laboratory of Evidence Science, Ministry of Education, China University of Political Science and Law, Beijing 100088, China;

²⁾Key Laboratory of Forensic Genetics, Ministry of Public Security, National Engineering Laboratory for Forensic Science, Institute of Forensic Science, Beijing 100038, China;

³⁾Jiangsu Key Laboratory of Phylogenomics and Comparative Genomics, School of Life Sciences, Jiangsu Normal University, Xuzhou 221116, China)

Graphical abstract



Abstract Objective To construct an analysis process for identity-by-descent (IBD) algorithm to predict distant relatives and evaluate the prediction accuracy. **Methods** 253 family samples were detected by using high-density whole genome single nucleotide polymorphism (SNP) chip. IBD algorithm was used to predict the genetic

* This work was supported by grants from The National Natural Science Foundation of China (82171870), the Ministry of Public Security Technology Research Program (2021JSZ15), and the Fundamental Research Funds for Institute of Forensic Science (2021JB004).

** Corresponding author.

LI Cai-Xia. Tel: 86-10-83752706, E-mail: licaixia@tsinghua.org.cn

ZHAO Dong. Tel: 86-15010055761, E-mail: dzfslm@163.com

Received: November 21, 2022 Accepted: April 11, 2023

relationship between pairs of individuals, and the prediction accuracy was evaluated. The number of SNPs was randomly reduced to evaluate the effect of SNP numbers on the accuracy of the algorithm prediction.

Results Among 1–7th kinship degree, the average confidence interval accuracy of IBD algorithm was 94.72%, and the prediction credibility was 99.77%. The false negative of IBD algorithm was found when kinship degree is 6 or higher. When the number of SNPs decreases, the prediction accuracy will decline to a certain extent.

Conclusion The IBD algorithm can accurately predict the genetic relationship within the seventh kinship degree, and it has important application value in population genetics, forensic genealogy inference and other fields.

Key words IBD algorithm, forensic genealogy inference, distant relatives prediction

DOI: 10.16476/j.pibb.2022.0538