PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2022,49(3):591~599
www.pibb.ac.cn

基于IBS算法预测亲缘关系准确性研究*

管珊珊[1,2] 张文杰[1,2] 魏以梁[3] 李鹰翔[4] 赵雯婷[2] 范 虹[1)**] 刘 京[2,5)**]

([1] 陕西师范大学计算机科学学院，西安 710119；[2] 公安部物证鉴定中心，北京 100038；[3] 江苏师范大学，徐州 221116；[4] 安澜智能（深圳）有限公司，深圳 510630；[5] 中国政法大学，证据科学教育部重点实验室，北京 100088）

摘要 **目的** 评估基于状态一致性（identity-by-state，IBS）算法预测个体间亲缘关系的准确性。**方法** 采用Illumina GSA芯片对253份样本进行全基因组检测，基于高密度单核苷酸多态性（single nucleotide polymorphism，SNP）数据计算两两个体间IBS共享统计量预测亲缘关系。通过不同条件参数筛选SNP，评估位点数对算法预测准确性的影响。**结果** 1~4级亲缘关系预测准确率高达99%，预测误差在1级以内且无假阳性。SNP数量减少对预测准确率无显著影响，即使在较低密度的SNP标记中，该算法也能获得较高的准确率。**结论** IBS算法是法医系谱推断的有效方法，且对于微量降解的法医现场检材具有很好的应用价值。

关键词 IBS算法，单核苷酸多态性，状态一致性，亲缘关系等级，法医遗传学，法医系谱推断
中图分类号 R89，Q812 **DOI**：10.16476/j.pibb.2021.0107

短串联重复序列（short tandem repeat，STR）基因座一直是司法领域中鉴定个体身份和亲缘关系的主要遗传标记，但由于使用的位点数目有限，常将单核苷酸多态性（single nucleotide polymorphism，SNP）遗传标记作为STR标记的补充。近年来，测序技术的进步发展带来了更密集的遗传标记集，由于SNP位点分布广泛、突变率低、相比STR等重复序列标记具有更高的遗传稳定性等特点，使得利用全基因组高密度SNP分型数据预测亲缘关系成为新的研究热点。法医系谱推断是指通过遗传谱系分析解决涉及司法实践中的身份识别问题，早在2005年，美国科学家Fitzpatrick提出此概念[1]。“金州杀手”案是第一宗使用法医系谱学技术破获的悬案[2]，该技术被誉为2018年度十大科学突破之一，此案告破后，警方利用该技术为200余例案件提供侦查线索[3-4]。2020年12月，国内利用法医系谱推断技术为14年前的一起命案积案锁定重点家系[5]，为案件侦破提供了直接线索。研究表明，高密度SNP技术结合传统STR技术将会成为法医DNA服务案件侦查和诉讼的新模式[6]。

个体间的共祖片段（identity-by-descent，IBD）长度算法或者等位基因频率估计的状态一致性（identity-by-state，IBS）共享统计量算法是目前预测亲缘关系的主要方式[7]。前者通过检测个体之间从一个共同祖先继承的相同DNA片段长度和数量，判断亲缘关系远近。该算法适用于已进行基因型定相的单倍型，需要较大的参考人群数据。并且对法医样本的质量非常敏感，当使用来自低质量DNA样本的少量SNP基因型时，很难实现可靠的IBD检测。基于等位基因频率估计IBS共享统计量预测亲缘关系的算法，在假设各标记间独立的情况下，通过估计整个样本中每个SNP的等位基因频率，计算基因组中共享的等位基因比例确定亲缘关系。该算法虽然只能准确预测1~4级内的亲缘关系，在5级以上的远亲关系中预测准确率低于IBD方法，但其受位点检出率影响较小。

* 国家科技资源共享服务平台计划（YCZYPT[2017]01-3），中央级公益性科研院所基本科研业务费专项资金（2019JB011，2021JB004），陕西省重点研发计划资助课题（2018SF-251），公安部“双十攻关”项目（2019SSJH0601）和首都科研领军人才培养工程（Z18110006318006）资助。

** 通讯联系人。
范虹 Tel：15929807273，E-mail：fanhong@snnu.edu.cn
刘京 Tel：18211057899，E-mail：biojing@yeah.net

收稿日期：2021-04-20，接受日期：2021-06-24

本文描述的IBS算法依赖于高密度SNP数据，通过计算每个SNP标记等位基因频率和IBS的共享等位基因数量估计两两个体之间的共享统计量，并转化为亲缘关系系数得出亲缘关系等级。该算法在项目组开发的亲缘关系预测系统（kinship prediction system version 1.0，KPS v1.0）[8]中实现，可准确预测4级以内的亲缘关系，并且能在几分钟内对数百万对个体进行关系推断[9]。

1　材料与方法

1.1　样本来源

采集中国中部地区5个家庭共253个汉族样本，其中包含4 184对1~7级亲缘关系（图1显示各等级数量分布，包括双胞胎（MZ）、亲子（PO）、全同胞（FS）、2级（2nd）、3级（3rd）、4级（4th）、5级（5th）、6级（6th）、7级（7th）亲缘关系），26 325对无亲缘关系（UN）。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物证鉴定中心伦理委员会审查（编号：2020-022）。

1.2　DNA提取与检测

所有样本均使用QIAamp DNA Midi试剂盒（QIAGEN公司，德国）提取DNA，使用NanoDrop 2000c超微量分光光度计（Thermo Scientific公司，美国）进行DNA定量和纯度检测。使用美国Illumina Infinium Global ScreeningArray（GSA）芯片进行全基因组SNP检测，获得约70万个常染色体SNP位点分型（安澜智能公司，中国）。

1.3　亲缘关系推断方法

使用亲缘关系预测系统KPS v1.0进行亲缘关系预测，该系统通过IBS算法估计的亲缘关系系数Φ和零IBD共享统计量π_0推断亲缘关系等级。具体来说，亲缘关系系数Φ_{ij}表示从个体i、j中随机抽取的两个等位基因来源于同一祖先的概率。

$$\Phi_{ij}=\frac{N_{Aa,Aa}-2N_{AA,aa}}{2N_{Aa}^{(i)}}+\frac{1}{2}-\frac{1}{4}\frac{N_{Aa}^{(i)}+N_{Aa}^{(j)}}{N_{Aa}^{(i)}} \quad (1)$$

其中$N_{AA,\,aa}$为个体i，j基因型都为纯合子的标记数，$N_{Aa,\,Aa}$为个体i、j基因型都为杂合子的标记数，$N_{Aa}^{(x)}$是个体x的基因型为杂合子的标记数。零IBD共享统计量π_0表示从个体i，j在一个SNP位点上共享同一祖先零个等位基因的概率。

$$\pi_0=\frac{N_{AA,aa}}{\sum_m 2p_m^2\left(1-p_m\right)^2} \quad (2)$$

其中p_m为标记m的估计等位基因频率个体。

表1是对Manichaikul等[9]文献中亲缘关系系数Φ和零IBD共享统计量π_0的推理标准的扩展。根据系统预测的所有个体间的亲缘关系系数与此表中亲缘关系系数的推理标准范围比对，可进行个体间亲缘关系等级推断。由于亲子与全同胞关系的亲缘关系系数范围一致，可使用零IBD共享统计量作进一步区分。

1.4　位点筛选

使用高密度SNP标记集进行亲缘关系预测时通常包含一定程度的冗余信息，故本研究分别使用连锁不平衡、最小等位基因频率对标记进行过滤，评估不同位点组合的预测准确率。并且考虑到在真实案例样本中，检材质量不一，可能会导致位点随机丢失，故本研究还进一步通过随机减少位点数模拟真实低质量样本，以检验该算法的适用性。

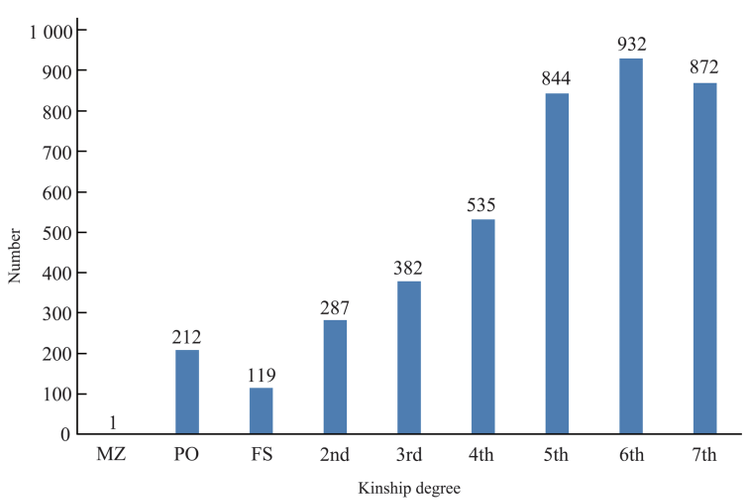

Fig. 1　The quantity distribution of each kinship degree among samples

Table 1　The criteria for inference of kinship

Relationship	Φ	Inference criteria	π_0	Inference criteria
MZ	$\frac{1}{2}$	$>\frac{1}{2^{3/2}}$	0	<0.1
PO	$\frac{1}{4}$	$(\frac{1}{2^{5/2}},\frac{1}{2^{3/2}})$	0	<0.1
FS	$\frac{1}{4}$	$(\frac{1}{2^{5/2}},\frac{1}{2^{3/2}})$	$\frac{1}{4}$	(0.1 , 0.365)
2nd	$\frac{1}{8}$	$(\frac{1}{2^{7/2}},\frac{1}{2^{5/2}})$	$\frac{1}{2}$	$(0.365 , 1-\frac{1}{2^{3/2}})$
3rd	$\frac{1}{16}$	$(\frac{1}{2^{9/2}},\frac{1}{2^{7/2}})$	$\frac{3}{4}$	$(1-\frac{1}{2^{3/2}},1-\frac{1}{2^{5/2}})$
4th	$\frac{1}{32}$	$(\frac{1}{2^{11/2}},\frac{1}{2^{9/2}})$	$\frac{7}{8}$	$(1-\frac{1}{2^{5/2}},1-\frac{1}{2^{7/2}})$
5th	$\frac{1}{64}$	$(\frac{1}{2^{13/2}},\frac{1}{2^{11/2}})$	$\frac{15}{16}$	$(1-\frac{1}{2^{7/2}},1-\frac{1}{2^{9/2}})$
6th	$\frac{1}{128}$	$(\frac{1}{2^{15/2}},\frac{1}{2^{13/2}})$	$\frac{31}{32}$	$(1-\frac{1}{2^{9/2}},1-\frac{1}{2^{11/2}})$
7th	$\frac{1}{256}$	$(\frac{1}{2^{17/2}},\frac{1}{2^{15/2}})$	$\frac{63}{64}$	$(1-\frac{1}{2^{11/2}},1-\frac{1}{2^{13/2}})$
UN	0	≤0	1	1

1.4.1　连锁不平衡

等位基因间的关联（连锁不平衡）会增加相邻遗传标记上等位基因共享的程度，从而高估个体间的亲缘关系程度，甚至将无关个体推断为有亲缘关系。考虑到全基因组SNP数据中存在大量连锁不平衡位点，本文使用PLINKv1.9软件[10]通过连锁不平衡的筛选标准R^2参数对原始数据进行位点过滤，以检验连锁不平衡是否对准确性产生影响。

1.4.2　最小等位基因频率

最小等位基因频率（minor allele frequencies，MAF）较低的位点对亲缘关系的信息贡献小，甚至增加假阳性，本文使用PLINK v1.9软件根据MAF值对原始数据进行位点过滤，去除冗余和非信息量标记位点，从而保留最大的信息量，以此评估位点信息量大小是否对预测结果造成影响。

1.4.3　随机筛选位点

由于法医学中常遇到陈旧、微量及降解检材等造成的位点检出率低的情况，为了探究SNP位点数量的减少对该算法预测效能影响，我们对位点进行随机的梯度下降筛选，将筛选的位点组合进行亲缘关系预测的结果与原始数据结果进行比较，检验不同密度SNP位点组合对预测准确性的影响，以及位点数量减少到何种程度，准确率会大幅下降。

2　结　　果

2.1　关系推断的准确性评估

使用KPS v1.0系统对253份测序数据进行亲缘关系计算，将所有个体间预测的亲缘关系等级与实际调查的亲缘关系进行比较，评估亲缘关系预测准确性。

表2中展示了253份样本数据进行亲缘关系预测的准确性，由于亲缘关系系数Φ在（0，0.002 76）范围的个体对亲缘关系无法确定，将不确定关系的样本对定义为7级以上或未知关系（>7th/UnK）。从表中可以看出，1级亲缘关系的预测准确率为100%，3级亲缘关系预测准确率为89.8%，随着亲

Table 2　The evaluation of the accuracy of genetic relationship prediction in 253 samples

Real kinship	Predict kinship											AC[1)]/%	CIA[2)]/%	FN[3)]/%	FP[4)]/%
	MZ	PO	FS	2nd	3rd	4th	5th	6th	7th	>7th /UnK	UN				
MZ	1											100	100	0	
PO		212										100	100	0	
FS			119									100	100	0	
2nd				281	6							97.9	100	0	
3rd				5	343	30	4					89.8	99.0	0	
4th					27	364	134	8	1		1	68.0	98.1	0.2	
5th					1	65	345	264	85	46	38	40.9	79.9	4.5	
6th						4	84	239	161	164	280	25.6	51.9	30.0	
7th							13	87	145	157	470	16.6	26.6	53.9	
UN							17	424	1 536	3 745	20 603				7.5

[1)]Accuracy (AC): the probability of correctly predicting two related individuals as the degree of their real kinship;[2)] Confidence interval accuracy (CIA): the probability that two related individuals is predicted to be within the range of 1st of their real kinship, and above 7th degree or unknown relationship is not taken into consideration;[3)] False negative (FN): the probability of predicting two related individuals as unrelated;[4)] False positive (FP): the probability of predicting two unrelated individuals as within 7th degree.

缘关系等级的增加，预测准确率也随之降低，4级开始出现假阴性，5级之后的亲缘关系准确率明显下降。

2.2　亲缘关系等级系数分布

基于调查的真实亲缘关系所估计的亲缘关系系数分布（图2），1~3级亲缘关系都比较清楚地分开，而4级以后的亲缘关系分布开始出现重叠，并且越远的关系与无关分布有更高程度的重叠。表3中为各亲缘关系等级系数的分布范围。

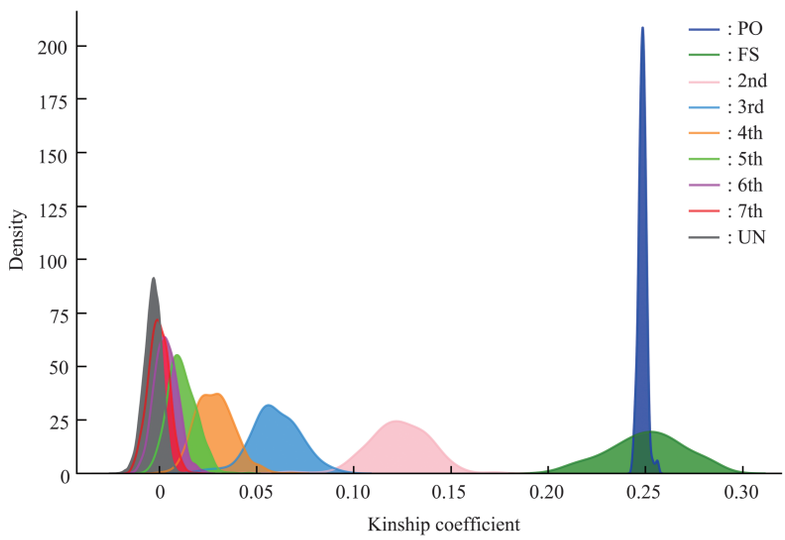

Fig. 2　The distribution map of kinship coefficient of each kinship degree

Table 3　The distribution range of kinship coefficient of each kinship degree

Real kinship	Kinship coefficient				
	Inference criteria	Minimum	Max	Mean	Standard deviation
MZ	>0.354	0.499 9	0.499 9	0.499 9	0
PO	[0.177, 0.354]	0.243 3	0.256 4	0.250 4	0.002
FS	[0.177, 0.354]	0.203 2	0.287 9	0.248 5	0.019 1
2nd	[0.088 4, 0.177)	0.045 3	0.173 9	0.122 8	0.016 4
3rd	[0.044 2, 0.088 4)	0.015 8	0.097	0.059 7	0.012 9
4th	[0.022 1, 0.044 2)	0	0.067 7	0.028 4	0.009 7
5th	[0.011 05, 0.022 1)	−0.007 8	0.045 7	0.011 5	0.007 2
6th	[0.005 52, 0.011 05)	−0.014 3	0.022 9	0.003 3	0.005 9
7th	[0.002 76, 0.005 52)	−0.014 2	0.018 9	−0.000 5	0.005 2
UN	≤0	−0.023 7	0.016 6	−0.003 5	0.004 5

2.3　不同SNP标记组合的关系预测

2.3.1　连锁不平衡

为研究连锁不平衡的对于亲缘关系预测的影响，本文根据连锁不平衡的度量参数R^2对位点进行过滤，使得保留的所有位点间的相关性都低于给定的R^2值。根据集合［0.1，0.125，0.15，0.175，0.2，0.225，0.25，0.275，0.3］中的值筛选位点，表4为不同R^2值筛选的位点组合预测准确性结果。图3中显示了不同位点组合在各亲缘关系等级的预测准确性分布。与原始数据预测准确性比较发现，R^2值越大，该算法的预测准确性越高，尤其对于4级以上的亲缘关系更为明显，例如5级的绝对准确率由40.9%升至56.8%，并且当R^2≥0.125时，消除了4级上唯一的一对假阴性结果。虽然筛选的位点在一定程度上提高了预测准确率，降低了总体的假阴性，但同时也增加了假阳性，并且在4级关系上出现假阳性结果。

Table 4　The predictive accuracy of locus combinations screened by different R^2−values

Real kinship	R^2=0.1（107 644 SNPs）				R^2=0.125（123 813 SNPs）				R^2=0.15（138 034 SNPs）			
	AC/%	CIA/%	FN/%	FP/%	AC/%	CIA/%	FN/%	FP/%	AC/%	CIA/%	FN/%	FP/%
MZ	100	100	0		100	100	0		100	100	0	
PO	100	100	0		100	100	0		100	100	0	
FS	100	100	0		100	100	0		100	100	0	
2nd	97.9	100	0		97.9	100	0		97.9	100	0	
3rd	90.6	99.5	0		91.4	99.5	0		92.1	99.5	0	
4th	72.0	98.9	0.2		73.8	99.1	0		73.1	99.1	0	
5th	49.3	84.0	2.0		51.3	85.5	2.0		50.9	86.4	1.7	
6th	34.5	68.1	16.7		35.4	72.6	14.1		35.2	72.2	14.6	
7th	19.8	37.7	36.7		17.4	39.6	32.5		21.0	45.9	28.8	
UN				10.3				11.1				11.8
Real kinship	R^2=0.175（151 186 SNPs）				R^2=0.2（163 577 SNPs）				R^2=0.225（175 188 SNPs）			
	AC/%	CIA/%	FN/%	FP/%	AC/%	CIA/%	FN/%	FP/%	AC/%	CIA/%	FN/%	FP/%
MZ	100	100	0		100	100	0		100	100	0	
PO	100	100	0		100	100	0		100	100	0	
FS	100	100	0		100	100	0		100	100	0	
2nd	97.9	100	0		97.9	100	0		97.9	100	0	
3rd	91.9	99.5	0		92.4	99.5	0		92.4	99.5	0	
4th	74.4	99.3	0		75.7	99.4	0		76.6	99.4	0	
5th	52.1	87.2	1.5		53.1	87.8	1.3		56.0	90.9	0.9	
6th	37.2	71.6	13.1		38.3	74.5	11.5		39.0	75.3	10.2	
7th	22.4	46.7	27.9		23.9	47.1	26.3		22.2	45.8	26.3	
UN				12.1				13.0				12.8
Real kinship	R^2=0.25（186 749 SNPs）				R^2=0.275（197 505 SNPs）				R^2=0.3（208 153 SNPs）			
	AC/%	CIA/%	FN/%	FP/%	AC/%	CIA/%	FN/%	FP/%	AC/%	CIA/%	FN/%	FP/%
MZ	100	100	0		100	100	0		100	100	0	
PO	100	100	0		100	100	0		100	100	0	
FS	100	100	0		100	100	0		100	100	0	
2nd	97.9	100	0		97.9	100	0		97.9	100	0	
3rd	92.4	99.5	0		92.1	99.5	0		91.9	99.7	0	
4th	75.7	99.4	0		76.6	99.6	0		76.8	99.6	0	
5th	55.1	91.6	1.1		55.0	91.8	0.9		56.8	91.9	0.6	
6th	38.3	75.2	10.0		38.4	76.1	9.5		39.7	76.1	8.7	
7th	23.2	46.6	28.3		24.1	46.3	26.9		26.1	47.4	26.0	
UN				12.6				12.8				12.2

2.3.2　最小等位基因频率

本文根据MAF值［0.000 1，0.01，0.05，0.1，0.2］对原始数据进行位点过滤，在筛选的结果数据集中，SNP标记数范围在222 770~514 962之间。使用过滤后的SNP位点组合进行亲缘关系预测（表5），SNP位点数量随MAF参数值增大而减少，预测准确率也随之降低。本文使用F检验分别将5组数据的准确性与原始数据的准确率进行计算，均得出F值在0.05的水平上无显著性差异（$F>F_{0.05}$）。因此可得，虽然不同的SNP位点组合对预测的结果会产生影响，但这种影响不显著。

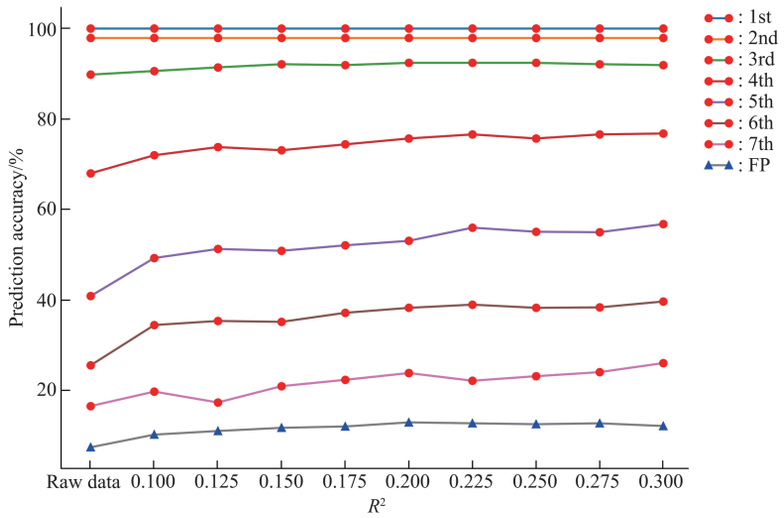

Fig. 3　The predictive accuracy of locus combinations screened by different R^2 values

Table 5　The predictive accuracy of locus combinations screened by different MAF−values

Real kinship	MAF=0.000 1 (514 962 SNPs) (The result is consistent with the original data)				MAF=0.01 (465 131 SNPs)				MAF=0.05 (374 980 SNPs)			
	AC/%	CIA/%	FN/%	FP/%	AC/%	CIA/%	FN/%	FP/%	AC/%	CIA/%	FN/%	FP/%
MZ	100	100	0		100	100	0		100	100	0	
PO	100	100	0		100	100	0		100	100	0	
FS	100	100	0		100	100	0		100	100	0	
2nd	97.9	100	0		97.9	100	0		97.9	100	0	
3rd	89.8	99.0	0		89.5	99.2	0		89.5	99.2	0	
4th	68.0	98.1	0.2		67.1	98.1	0		65.0	97.4	0	
5th	40.9	79.9	4.5		41.0	81.0	4.3		38.9	77.5	5.0	
6th	25.6	51.9	30.0		26.7	52.2	30.2		25.1	52.7	31.5	
7th	16.6	26.6	53.9		16.6	27.3	53.7		14.6	25.6	54.8	
UN				7.5				7.4				7.1

Real kinship	MAF=0.1 (312 985 SNPs)				MAF=0.2 (222 770 SNPs)			
	AC/%	CIA/%	FN/%	FP/%	AC/%	CIA/%	FN/%	FP/%
MZ	100	100	0		100	100	0	
PO	100	100	0		100	100	0	
FS	100	100	0		100	100	0	
2nd	97.9	100	0		97.9	100	0	
3rd	89.5	99.2	0		88.7	99.2	0	
4th	64.9	96.8	0		63.7	95.9	0	
5th	37.9	77.3	5.1		36.6	72.7	9.2	
6th	25.0	52.3	31.4		22.8	49.5	34.5	
7th	15.8	26.7	54.8		11.9	24.1	58.0	
UN				7.3				9.7

2.3.3　随机筛选位点

案件现场的生物检材受时间和环境等因素影响，DNA会发生降解，从而降低样本检出率。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699 537个SNP位点中，随机筛选40万、5万、1万和5 000各10组数据，使用IBS算法预测亲缘关系，预测准确性以平均值和标准差反映。表6结果显示，准确性随位点数量的减少而轻微降低，对3级内的亲缘关系准确性影响很小。但需要注意的是，当位点减少到5万个SNP时，4级亲缘关系预测开始出现假阳性，位点数量下降至1万时，少量无关样本被预测为3级。

Table 6　The predictive accuracy of random screening of different number of locus combinations

Real kinship	400 000 SNPs				50 000 SNPs			
	AC/%	CIA/%	FN/%	FP/%	AC/%	CIA/%	FN/%	FP/%
MZ	100	100	0		100	100	0	
PO	100	100	0		100	100	0	
FS	100	100	0		100	100	0	
2nd	97.9	99.9±0.13	0		97.3±0.30	99.3±1.09	0	
3rd	89.3±0.68	99.3±0.21	0		85.4±1.71	99.1±0.31	0	
4th	66.8±0.99	97.8±0.46	0.1±0.14		58.9±3.06	93.0±1.34	0.5±0.30	
5th	40.3±0.86	79.0±1.24	4.6±0.59		35.1±2.95	68.4±3.78	14.2±4.63	
6th	25.1±1.48	52.0±1.56	30.8±1.79		20.0±1.91	44.8±3.92	41.6±4.43	
7th	15.0±1.27	26.3±1.58	54.6±1.39		11.2±1.30	23.4±4.53	56.6±7.06	
UN				8.6±0.48				16.7±2.37
Real kinship	10 000 SNPs				5 000 SNPs			
	AC/%	CIA/%	FN/%	FP/%	AC/%	CIA/%	FN/%	FP/%
MZ	100	100	0		100	100	0	
PO	100	100	0		100	100	0	
FS	100	100	0		100	100	0	
2nd	95.3±1.32	99.6±0.28	0		91.5±1.17	99.6±0.42	0	
3rd	74.6±3.68	97.0±0.73	0.5±0.61		63.9±4.32	92.7±1.97	1.1±0.67	
4th	47.0±4.28	82.2±5.01	4.96±2.21		35.7±2.56	71.6±4.41	12.9±2.86	
5th	26.7±3.76	57.3±7.40	27.8±6.10		19.5±2.25	48.7±4.49	38.8±4.67	
6th	13.4±1.51	37.4±5.31	49.4±7.23		10.1±0.90	29.7±3.79	55.3±4.70	
7th	7.0±1.34	18.6±3.07	61.0±8.10		4.5±0.96	12.9±1.88	63.2±4.68	
UN				23.8±3.21				27.8±3.74

3　讨　　论

在法医遗传学领域，利用密集SNP标记数据预测亲缘关系的应用研究受到越来越多的关注，但目前缺乏针对中国人群的系统研究，包括从大规模SNP基因型数据集中筛选适合中国人群亲缘关系预测的位点组合，建立预测算法并对算法相关参数进行研究分析。本文中描述的IBS算法基于全基因组SNP数据进行关系推理，其框架核心是将一对个体之间的遗传距离作为其等位基因频率和亲属关系系数的函数进行建模，从而预测亲缘关系等级。该算法能够快速准确预测4级以内的亲缘关系，平均准确率可达99%以上。与IBD算法相比，此算法不需要特殊的计算资源，能在几分钟内对数百万对个体进行关系推断[9]。

本研究采用高密度SNP芯片对253份汉族样本进行检测，采用项目组前期开发的KPS v1.0系统进行亲缘关系预测，该系统将IBS算法的整体分析流程进行集成，实现了程序自动化。预测结果（表2）表明，IBS在4级以内的准确率极高，在1级误差内，4级预测准确率高达98.1%。

基于253份样本真实亲缘关系的亲缘关系系数分布（图2）显示，1~3级关系能明显分离开来，而4级以上的亲缘关系系数会出现重叠，最远的7

级关系与无关关系重叠最大。该结果与预期一致，由于亲缘关系越远的个体间共享的等位基因数量相对较少使得亲缘关系系数降低，且从表3中可以看出，4级以上的亲缘关系系数的均值开始与对应的推理标准值出现明显偏差，因此该方法对于此类关系难以准确区分。

本研究还进一步探讨了连锁不平衡、最小等位基因频率以及位点数量对该算法的影响，以筛选适合中国人群的系谱分析位点组合。首先考虑到密集SNP遗传标记数据中存在大量连锁位点、冗余信息等现象，可能对预测结果产生影响，研究中通过连锁不平衡度量参数R^2值对原始数据进行位点筛选，与原始数据预测准确性比较发现，随着R^2值越大，该算法的预测准确性越高，尤其对于4级以上的亲缘关系更为明显（表4和图3）。虽然筛选的位点在一定程度上提高了预测准确率，降低了总体的假阴性，但在4级关系上出现假阳性结果。因此，使用此参数时，预测结果中的假阳性和假阴性率需要均衡。其次，MAF使用近似0值过滤了不提供信息的位点，预测结果（表5）与原始结果一致，并减少了计算时间（本文中未体现）。其他参数值的预测结果（表5）显示，预测准确性随MAF值的增加而轻微降低，该结果很有可能由于位点数减少的影响。虽然最小等位基因频率对预测的结果会产生影响，但这种影响不显著。由此可知，该算法对此参数并不敏感。最后，在法医工作中，标记密度不一定能得到保证，比如脱落细胞、腐败组织等微量降解的案件检材。本研究进一步通过随机筛选位点数量模拟低质量DNA样本的少量SNP位点进行亲缘关系预测的准确性。结果表明（表6），预测准确性随位点数量的减少而降低，但影响较小。并且所有结果显示，位点数量对近亲缘关系影响更小，比如亲子、全同胞以及2级关系。但值得注意的是，使用5 000个SNP位点进行计算时，3级关系的预测准确性能达到92.7%，误差在1级内，与Kling等[11]研究结果（至少需要5.6万个SNP来确定一代堂表兄妹（3级关系））相比，该算法预测效能很好，在较低密度SNP中也能以较高准确率预测4级以内关系。

4 结　论

本文探索研究了基于高密度SNP数据利用IBS算法进行亲缘关系预测的可行性，研究结果表明，该算法在4级以内亲缘关系的预测效能很好，并且此算法受SNP位点数量减少的影响较小，对于陈旧降解等低质量检材，也能保持较高的准确性。因此该方法可辅助物证鉴定工作，为刑事犯罪、灾难受害者身份识别（disaster victims identification，DVI）、冷案积案等疑难案件的侦破提供科技支撑。

参　考　文　献

[1] Fitzpatrick C . Forensic Genealogy. Fountain Valley, CA: Rice Book Press, 2005

[2] Phillips C. The Golden State Killer investigation and the nascent field of forensic genealogy. Forensic Sci Int Genet, 2018, **36**:186-188

[3] 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

[4] Greytak E M, Moore C, Armentrout S L. Genetic genealogy for cold case and active investigations. Forensic Sci Int, 2019, **299**:103-113

[5] 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 Science and Technology, 2021,**46**(6):652-656

[6] Ge J, Budowle B. How many familial relationship testing results could be wrong?. PLoS Genet, 2020, **16**(8):e1008929

[7] 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

[8] 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 Science Research, 2020, **24**(3): 229-233+258

[9] Manichaikul A, Mychaleckyj J C, Rich S S, *et al*. Robust relationship inference in genome-wide association studies. Bioinformatics, 2010, **26**(22):67-73

[10] Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, 2007, **81**(3):559-575

[11] Kling D, Phillips C, Kennett D, *et al*. Investigative genetic genealogy: current methods, knowledge and practice. Forensic Sci Int Genet, 2021, **52**:102474

Accuracy Research on The Kinship Relationship Prediction by IBS Algorithm*

GUAN Shan-Shan[1,2], ZHANG Wen-Jie[1,2], WEI Yi-Liang[3], LI Ying-Xiang[4], ZHAO Wen-Ting[2], FAN Hong[1]**, LIU Jing[2,5]**

([1]*School of Computer Science, Shaanxi Normal University, Xi'an* 710119, China;
[2]*Institute of Forensic Science, Ministry of Public Security (MPS), Beijing* 100038, China;
[3]*Jiangsu Normal University, Xuzhou* 221116, China;
[4]*AnLan AI (Shenzhen) Ltd., Shenzhen* 510630, China;
[5]*Ministry of Education's Key Laboratory of Evidence Science, China University of Political Science and Law, Beijing* 100088, China)

Abstract **Objective** To evaluate the accuracy of predicting the kinship relationship between individuals based on the identity-by-state (IBS) algorithm. **Methods** The Illumina GSA chip was used to perform whole-genome detection on 253 samples. Based on high-density single nucleotide polymorphism (SNP) data, the IBS sharing statistics between two individuals was calculated to predict the kinship relationship. Filtering SNP by different conditional parameters to evaluate the influence of the number of sites on the accuracy of the algorithm's prediction. **Results** The prediction accuracy rate of 1st-4th degree of relatives proved to be as high as 99%, with a paired difference of 1st degree and no false positive. Decrease in the number of SNPs has no significant impact on the accuracy of prediction, and the algorithm still achieves a higher accuracy rate even in the lower density of SNP markers. **Conclusion** The IBS algorithm provides an effective method for forensic genealogy inference, which has good application value for forensic on-site inspection materials with trace degradation.

Key words IBS algorithm, single nucleotide polymorphism, identity-by-state, degree of relatedness, forensic genetics, forensic genealogy inference
DOI：10.16476/j.pibb.2021.0107

* This work was supported by grants from the National Science and Technology Resources Sharing Service Platform Project (YCZYPT[2017]01-3), the Fundamental Research Funds for Institute of Forensic Science (2019JB011, 2021JB004), the Key Research and Development Program of Shaanxi Province(2018SF-251), the Ministry of Public Security Double Ten Key Project (2019SSJH0601) and Beijing Leading Talent Project (Z18110006318006).
** Corresponding author.
FAN Hong. Tel：86-15929807273，E-mail：fanhong@snnu.edu.cn
LIU Jing. Tel：86-18211057899，E-mail：biojing@yeah.net
Received：April 20，2021 Accepted：June 24，2021