



基于全基因组测序的法医单核苷酸多态性 系谱推断研究*

谢全^{1,2)} 桂虚^{1,2,3)} 赵雯婷²⁾ 方之晓²⁾ 徐景怡²⁾ 刘京²⁾ 李彩霞^{1,2)**}

⁽¹⁾ 中国人民公安大学侦查学院, 北京 100038;

²⁾ 公安部物证鉴定中心, 现场物证溯源技术国家工程实验室, 北京市现场物证检验工程技术研究中心, 北京 100038;

³⁾ 四川省达州市公安局达川区分局, 达州 635000)

摘要 目的 通过全基因组测序 (whole genome sequencing, WGS) 获得高密度单核苷酸多态性 (single nucleotide polymorphism, SNP) 分型数据, 评估分型准确性, 研究建立 WGS 数据用于法医 SNP 系谱推断的方法。方法 通过华大 MGISEQ-200RS 测序平台对样本进行深度为 30× 的 WGS, 从测序数据中提取 Wega GSA 芯片中的 645 199 个常染色体 SNP 位点, 质控过滤后运用 IBS/IBD 算法计算预测亲缘关系, 并对样本的族群来源进行分析。结果 从测序数据中提取的 SNP 分型与 Wega GSA 芯片分型的一致率大于 99.62%。测序获得的 SNP 数据使用 IBS 算法可预测 1~4 级亲缘关系, 4 级亲缘预测置信区间准确性达 100%, 使用 IBD 算法可预测 1~7 级亲缘关系, 7 级亲缘预测为有亲缘关系的准确性达 100%, 通过高精度 WGS 数据获取的 SNP 系谱推断能力与芯片预测结果无显著差异。同时, WGS 数据用于族群推断与调查结果一致。结论 WGS 技术可应用于法医 SNP 系谱推断, 为案件侦破提供线索。

关键词 单核苷酸多态性, 全基因组测序, 系谱推断, 亲缘关系

中图分类号 RD89, D919

DOI: 10.16476/j.pibb.2022.0506

单核苷酸多态性 (single nucleotide polymorphism, SNP) 在人基因组中分布广泛, 突变率低, 遗传稳定性高, 是法医遗传领域的第三代遗传标记, 常用于族群地域推断、体貌表型预测、亲缘关系分析等^[1]。其中, SNP 在亲缘分析尤其是远距离亲缘分析方面具有独特的优势。第二代遗传标记短串联重复序列 (short tandem repeat, STR) 标记通常用于亲子鉴定、全同胞鉴定, 但是对二级以上亲缘关系难以准确鉴定^[2]。2018 年, 美国“金州杀手案”的破获就是通过高密度 SNP 预测出嫌疑人的 7 级远亲关系, 进而找到嫌疑人, 该技术被称为法医 SNP 系谱推断技术, 并被《科学》(Science) 杂志评为年度全球十大科学突破之一^[3]。

法医 SNP 系谱推断技术迅速成为近年来法医领域的研究热点, 并为数百起冷案积案的侦破提供了关键线索^[4]。本项目组也建立了相关技术并在命案积案侦破中发挥作用^[5]。法医 SNP 系谱推断技术是基于高密度 SNP 数据, 通过个体间的共祖片段 (identity by descent, IBD) 长度算法、状态一致性 (identity by state, IBS) 共享统计量等算法

预测个体间的亲缘关系等级^[6]。获得高密度 SNP 分型数据的方式主要为全基因组 SNP 芯片技术和全基因组测序 (whole genome sequencing, WGS) 技术。其中, 由于全基因组 SNP 芯片的价格优势, 成为目前法医 SNP 系谱推断最常用的数据获取方式^[7]。SNP 芯片是在一块基片表面固定序列已知的靶核苷酸的探针, 用于特定 SNP 位点组合的检测分型。国内外常用芯片包括 GSA 芯片 (65 万 SNP)、CytoSNP 芯片 (85 万 SNP)、Affymetrix 定制芯片 (20 万~260 万 SNP) 等, 均为 Illumina、Affymetrix 等国外产品。而且, SNP 芯片技术对 DNA 质量的要求较高, 一般需 200 ng 以上, 250 pg 时 7 级亲缘预测准确性由 99.9% 降至 62.6%, 对血样、精斑、唾液等样本检测效果好, 但是微量降解等现场常见生物检材的检测能力差, 模拟降解

* 国家自然科学基金 (82171870) 和公安部技术研究计划 (2021JSZ15) 资助项目。

** 通讯联系人。

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

收稿日期: 2022-10-26, 接受日期: 2023-03-01

检材平均DNA片段大小为150 bp时,其准确率降至0^[7]。WGS技术又称为大规模并行测序,区别于传统的Sanger(双脱氧法)测序,能够一次并行对大量核酸分子进行序列测定的技术。WGS检测方案的一个优势在于,DNA片段化处理是文库构建的第一步,增加了从降解的法医样本中获得SNP分型结果的可能性^[8]。近期有研究发现,WGS技术在微量降解等法医现场检材的检测能力优于全基因组SNP芯片技术^[9]。

本项目组前期已构建起基于全基因组SNP芯片技术的法医SNP系谱推断技^[5-6],为建立国产平台检测能力,并进一步提升法医现场检材检测能力,本文使用华大MGISEQ-200RS测序平台进行高深度的WGS,旨在探索从高深度WGS数据中获得准确的SNP分型,并进行法医SNP系谱推断的可行性和准确性研究。

1 材料与方法

1.1 样本来源及DNA提取

采集本实验室来自不同家系的静脉血样本A、B、C,标准品使用NA12878(具有北欧和西欧血统的犹他州女性)。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:2021-006)。静脉血均使用QIAamp Blood Midi Kit试剂盒(QIAGEN公司,德国)提取DNA,使用Qubit™ dsDNA HS Assay Kit试剂盒在Invitrogen Qubit 4(Thermo Fisher公司,美国)上进行DNA定量,使用JY-SPDT电泳仪(君意东方公司,中国)判断DNA完整度。

1.2 文库制备和测序

质检合格的DNA,使用MGIEasy酶切DNA文库制备试剂套装(华大智造公司,中国),对

200 ng起始样本进行酶切文库构建。利用MGISEQ-200RS高通量测序试剂盒将单链环状文库制备成DNA纳米球(DNA nano ball, DNB),采用PE150读长测序策略,进行平均深度30×WGS。所有样本均使用MGISEQ-200RS测序仪进行测序。

1.3 测序数据处理

利用FastQC软件对测序仪下机数据的fastq文件进行质检,使用Fastp软件去除低质量片段和接头序列,生成Clean Fastq文件。使用BWA工具,将Clean Fastq数据比对到人类参考基因组(hg19)上生成bam文件。用Sort工具,将bam文件按照染色体组型进行排序,生成sort.bam文件。使用GATK软件Markduplicate工具标记重复的读段,选取唯一对比读段进行后续分析。通过GATK软件中BQSR工具用于矫正测序碱基的真实质量值,设置参考SNP数据集作为真实变异集,这些数据集之外的位点,若测序数据与参考基因组存在差异,则认为是错误,以此来判定并矫正测序碱基的质量值。使用GATK的HaplotypeCaller工具,进行二倍体基因组的变异检测,识别出潜在变异区域。

使用Bcftools软件从全基因组vcf文件中提取感兴趣的位点,利用Python 3脚本过滤低质量位点进行质控,并删除Indel。使用Plink软件将vcf文件转换为二进制格式,与参考样本合并进行法医SNP系谱推断。

1.4 家系参考样本

家系参考样本来自项目组志愿者共计3个家系,亲缘关系对应等级数量分布如表1,包含1~8级亲缘,234份样本^[6],使用Wegene GSA芯片(安澜智能公司,中国)获得645 199个常染色体SNP位点分型,参考刘京等^[10]法医系谱亲缘关系等级进行划分。族群推断参考样本来自公开的千人基因组数据库。

Table 1 The quantity distribution of each kinship degree among samples

Pedigree	1st	2nd	3rd	4th	5th	6th	7th	8th	UN	Total
A	4	5	0	0	0	0	0	0	11	20
B	2	6	3	2	1	1	0	0	4	19
C	3	3	4	17	21	29	11	15	92	195
Total	9	14	7	19	22	30	11	15	107	234

1.5 系谱推断

通过本项目组开发的亲缘关系预测系统,分别使用IBS和IBD算法对样本进行亲缘关系预测^[5-6]。

IBS算法通过计算亲缘关系系数 ϕ 和零IBD共

享统计量 π_0 推断两个体间亲缘关系等级。零IBD共享统计量 π_0 代表两个体在一个SNP位点上共享同一祖先零个等位基因的概率,亲缘关系系数 ϕ 表示从两个体中随机抽取的两个等位基因来源于同一祖

先的概率。根据系统预测的所有个体间 $\varnothing$ 和 π_0 与推断等级标准范围对比, 可进行个体间亲缘关系等级推断^[6]。

IBD是指由于在遗传过程中, 染色体会发生重组, 两个有亲缘关系个体的基因组中有一段DNA来源于同一个祖先, 通常用重组频率的测量单位厘摩 (centimorgan, cM) 衡量, IBD片段越长, 说明亲缘关系越近^[11]。预测系统基于背景人群参考数据进行同源染色体分离, 计算两两个体之间共祖片段IBD的cM总长度, 进而推断个体间亲缘关系等级^[12]。

法医系谱推断效能通过绝对准确率 (accuracy, AC)、置信区间准确率 (confidence interval accuracy, CIA)、假阴性率 (false negative, FN) 和假阳性率 (false positive, FP) 等指标评估。绝对准确率为预测亲缘等级与调查等级一致的关系对数占关系对数的比重, 置信区间准确率为预测等级在调查等级的上下一级浮动的关系对数占等级数的比重, 假阴性率为调查有亲缘关系被预测为无关的对数占亲缘对数的比重, 假阳性率为调查无关被预测为有关的对数占亲缘对数的比重。绝对准确率和置信区间准确率越高, 假阴性率和假阳性率越低, 系谱推断体系越准确。

1.6 族群分析

以千人基因组数据库^[13]样本为参考数据集, 利用全基因组SNP分型, 使用Plink v1.9软件进行PCA分析, R软件绘制PCA图。提取实验室前期研究的27AISNP体系位点, 使用DAA软件进行欧洲、非洲、东亚三大人群祖先来源推断^[14]。

2 结 果

2.1 测序质量评估

从3份中国北方人群的静脉血和标准品NA12878中获得的4份DNA, 分为4次样本制备和测序, 其中静脉血样本用于亲缘关系推断, 标准品用于测序准确性评估。

测序芯片FCL PE150属于单线型泳道, 满载输出约150 G fastq测序数据。表2中列出了芯片泳道的测序指标, 其中每条泳道平均产生 (642.65±25.63) 百万次读数, 平均芯片利用率为73.69%, 平均Q30比率为90.50%, 过滤后有效点占比ESR的均值为73.69%。MGISEQ-200RS平台与已发表的BGISEQ-500RS平台^[15]和MGISEQ-2000RS平台^[16]测序质量对比, 平台获得相似的Q30值。测序平均Q30为90.50%, 能够进行准确的基因组测序, 是后续一致性和推断准确性分析的数据基础。

Table 2 Run metrics for 4 runs in MGISEQ-200RS sequencing platform

Parameters	RUN1	RUN2	RUN3	RUN4	Average value
Chip productivity ¹⁾ /%	73.88	71.31	73.02	76.53	73.69±2.84
Total reads/M	644.41	621.15	636.75	668.28	642.65±25.63
Q30 ²⁾ /%	89.72	90.28	90.77	91.24	90.50±0.78
Split rate ³⁾ /%	99.04	98.90	99.02	99.15	99.03±0.13
Lag1 ⁴⁾ /%	0.08	0.08	0.08	0.08	0.08
Lag2 ⁴⁾ /%	0.10	0.09	0.10	0.09	0.095±0.005
Runon1 ⁴⁾ /%	0.07	0.07	0.07	0.07	0.07
Runon2 ⁴⁾ /%	0.06	0.07	0.07	0.07	0.068±0.008
ESR ⁵⁾ /%	73.88	71.31	73.02	76.53	73.69±2.84

¹⁾Chip productivity:the ratio of the number of fields of view (FOVs) that can normally output data to the total number of lane FOVs, multiplied by the ratio of effective reads;²⁾Q30%: basecall results with a base error rate of less than 0.001 (accuracy higher than 99.9%);³⁾Split rate%: total reads split into the correct probability of the corresponding barcode; ⁴⁾Lag/Runon: similar to phasing/prephasing in Illumina sequencing; ⁵⁾ESR%: effective spot rate, proportion of effective spots after filtering.

4份测序样本总计获得2 612 M reads, 比对到参考基因组hg19的reads为2 608 M, 比对率为99.8%。平均比对质量为54.18, 去除重复序列后样本深度在30×以上, 基因组1×平均覆盖率为

91.60%。GC含量均在40%左右, 与人类的基因组GC含量一致, 说明测序过程碱基的偏好性低, 且读段覆盖更均匀^[17]。所有样本的详细测序和SNP覆盖指标见表3。

Table 3 Statistical table of bamQC parameters

Parameters	A	B	C	NA12878
Number of reads/M	634	642	634	702
Mapping rate/%	99.91	99.91	99.90	99.65
Median insert size	296	294	295	265
Mean mapping quality	54.49	54.56	54.52	53.14
Coverage rate/%	91.56	92.22	92.26	90.34
GC percentage/%	40.30	39.59	40.18	39.51
Duplication rate/%	17.84	19.25	16.78	31.25
Depth (×)	30.4	30.7	30.3	33.6

全基因组覆盖率随深度阈值的变化如图1所示,在1×至5×深度区间,随着深度阈值的提升覆

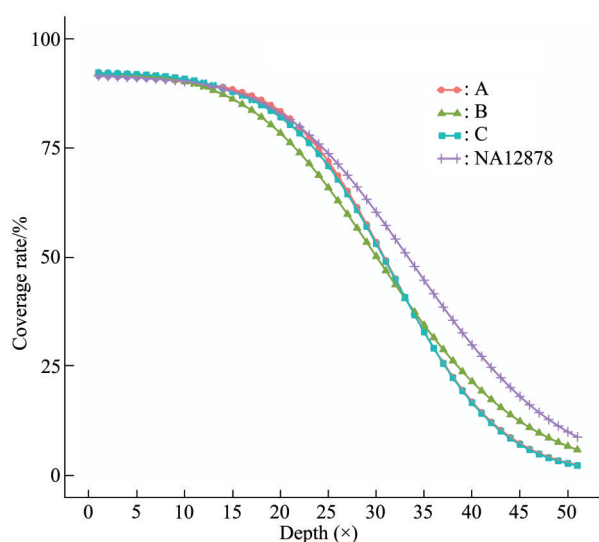


Fig. 1 Coverage rate of various samples in different depths

盖率降低缓慢,但在大于5×区域下降快速。覆盖率随深度增加的变化趋势在平均深度处达到最大。以平均深度为阈值,各样本覆盖率接近50%,说明各位点的测序深度集中在平均测序深度附近,但不可避免存在一部分位点深度较低及未检测到的情况。不同深度与覆盖率的关系反应了建库片段回收的随机性与高通量测序的随机性。随着平均测序深度的增加,数据量越大,冗余越多,随机性会相应的降低。因此,平均深度30×的WGS,可获得大量的系谱推断SNP位点。

从4份测序数据中提取Wegene GSA常染色体645 199个SNP位点,获得1×及以上深度认为该位点被成功提取。NA12878、A、B、C样本成功提取的SNP位点数目分别为:644 897、645 029、644 882、645 014个。对SNP位点进行测序深度统计(图2),平均测序深度处SNP位点数量达到峰值。

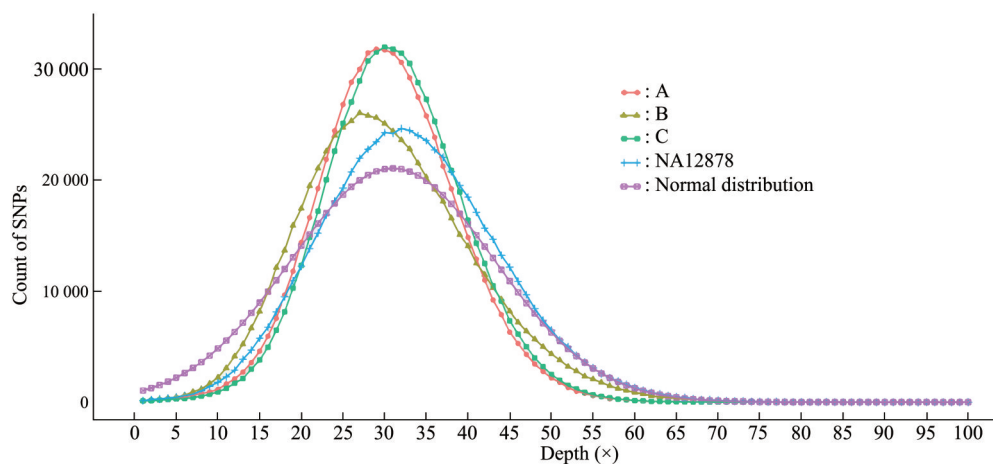


Fig. 2 Quantity distribution of SNP at different depths

2.2 SNP位点分型一致性评估

从千人基因组数据库NA12878的SNP位点分型文件提取Wegene GSA常染色体645 199个位点, 共获得621 128个SNP位点。从33.6×NA12878标准品测序数据中提取该621 128个SNP位点, 用于SNP分型准确性评估, 成功提取出621 127个SNP位点。

33.6×未质控标准品SNP位点与1000G数据库^[13]中NA12878的SNP位点进行一致性对比, 一致率达99.62%, 分型不一致位点2 365个。未质控

标准品分型为杂合子而千人数据分型为纯合子的SNP位点表现出位点分型的假阳性, 假阳性位点589个。未质控标准品测序分型为纯合子而千人数据分型为杂合子的位点表现为杂合性丢失, 为假阴性位点, 共1 348个。425个SNP位点表现为都为纯合位点且不一致, 3个SNP位点表现为杂合不一致。分别统计假阳性、假阴性、纯合不一致和杂合不一致位点测序深度(depth, DP)及测序质量值(genotype quality, GQ)(图3)。

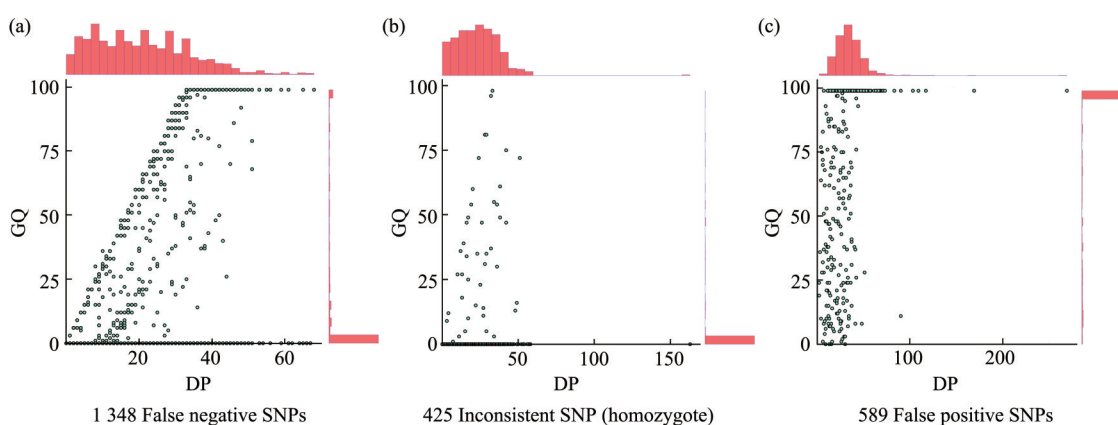


Fig. 3 Na12878 and 1000G inconsistent SNP mass distribution

不一致位点中, 假阴性与纯合不一致位点主要集中于测序质量值GQ较低的区域, 为确保在较高检出率的前提下获得高一致率, 以DP大于等于4, GQ大于等于10为阈值, 对测序样本进行质控。质控样本的假阴性与纯合不一致位点数减少明显, 假阳性位点数量却没有得到明显改善(图4)。假阳性位点存在的问题主要原因分为3点: a. 文库构建过程中PCR扩增错误; b. 测序仪对位点荧光信号的错误判读; c. 千人基因组数据该位点测序深度较低导致的SNP位点杂合性丢失。因此, 提高质控标准难以解决假阳性问题, 并且可能造成大量分型正确位点的丢失。质控后的杂合不一致位点并没有被过滤掉, 33.6×NA12878的rs6686181、rs72781591、rs9270230分型为AT、GT、CT, 位点测序深度分别为11×、26×和53×, 1000G数据中rs6686181、rs72781591、rs9270230分型AG、AG、CA。rs72781591位点的26×测序中18次检测到碱

基G, 2次检测到碱基A, 6次检测到碱基T, rs9270230位点出现了与rs72781591一样的情况, 23次检测到碱基C, 10次检测到碱基A, 20次检测到碱基T。MGISEQ-200RS平台采用红绿双色荧光激发系统, 碱基A能被红绿双色荧光激发, 碱基T仅被绿光激发, 碱基C仅被红光激发, 碱基G不能被激发, 可能为MGISEQ-200RS的双色荧光激发系统在判读该位点时绿光信号偏强, 从而导致碱基A、G被错误认为碱基T。

质控过滤后, 纯合不一致和假阴性位点明显减少, 不一致位点减少超过50%, 一致率得到提升, 但同时导致检出率的降低(表4)。提取率为深度大于等于1×的SNP位点个数在Wegene GSA芯片常染色SNP位点集合的比重, 检出率指通过深度大于等于4×、QG值大于等于10指控条件的SNP位点个数在Wegene GSA芯片常染色SNP位点集合的比重。

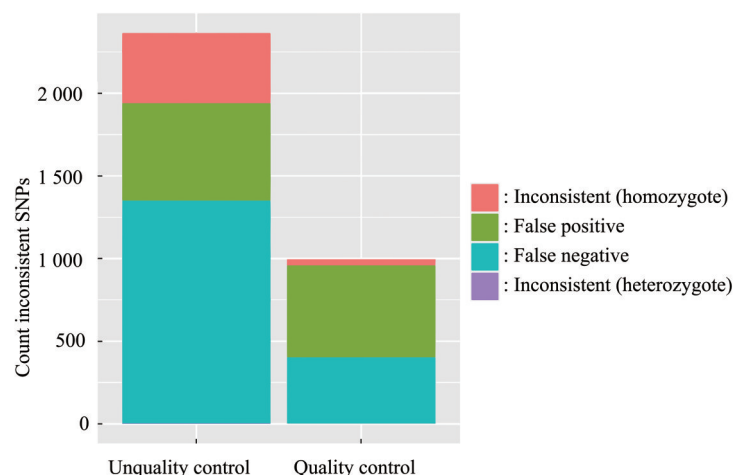


Fig. 4 Comparison of inconsistent loci in NA12878 after quality control

Table 4 Consistency comparison parameter statistics

Sample	Call rate/%	Overlap SNP	Consistent SNP	Consistent rate/%	Inconsistent SNP	False negative	False positive	Inconsistent (homozygote)	Inconsistent (heterozygote)
NA12878 (unquality control)	99.85	621 127	618 762	99.62	2 365	1 348	589	425	3
NA12878 (quality control)	99.21	617 245	616 249	99.84	996	399	557	37	3

从测序样本中提取 Wegene GSA 常染色体位点, 各样本进行GQ为10的质控过滤, 质控后检出率和一致率如表5。由于A、B样本仅有 Wegene CGA 芯片分型数据做为对比。A、B检出率为99.34%和99.28%, 提取 Wegene GSA 和 Wegene CGA 芯片交集位点500 753个, 一致率分别为99.84%和99.90%。C检出率为99.36%, 与其 Wegene GSA 芯片分型进行对比, C共有一致位点636 587个, 一致率为99.62%。

Table 5 Samples call rate after quality control

Sample	A	B	C
Call SNP	640 970	640 577	641 039
Overlap SNP	497 030	496 407	639 059
Consistent SNP	496 254	495 917	636 587
Call rate/%	99.34	99.28	99.36
Consistent rate/%	99.84	99.90	99.62

2.3 SNP位点在染色体上分布情况

质控后, 样本A、B、C高深度测序数据均获得64万以上的位点分型数据, 位点在染色体上的分布如图5, 结果显示各样本与GSA芯片位点分布完全一致, 较为均匀的分布于22对常染色体上, 局部区域特征上端粒区域SNP分布略高于染色体中部区域, 其中所有样本6号染色体28Mb HLA区域SNP分布密集。位点在染色体上分布情况, 说明WGS不会导致染色体上某区域特异性检测问题, 并且满足高密度SNP系谱推断技术位点在人类基因组中均匀分布的要求。

2.4 亲缘关系预测

A、B、C均为平均深度30×测序数据, 与家系参考样本Wegene GSA芯片数据合并, 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。

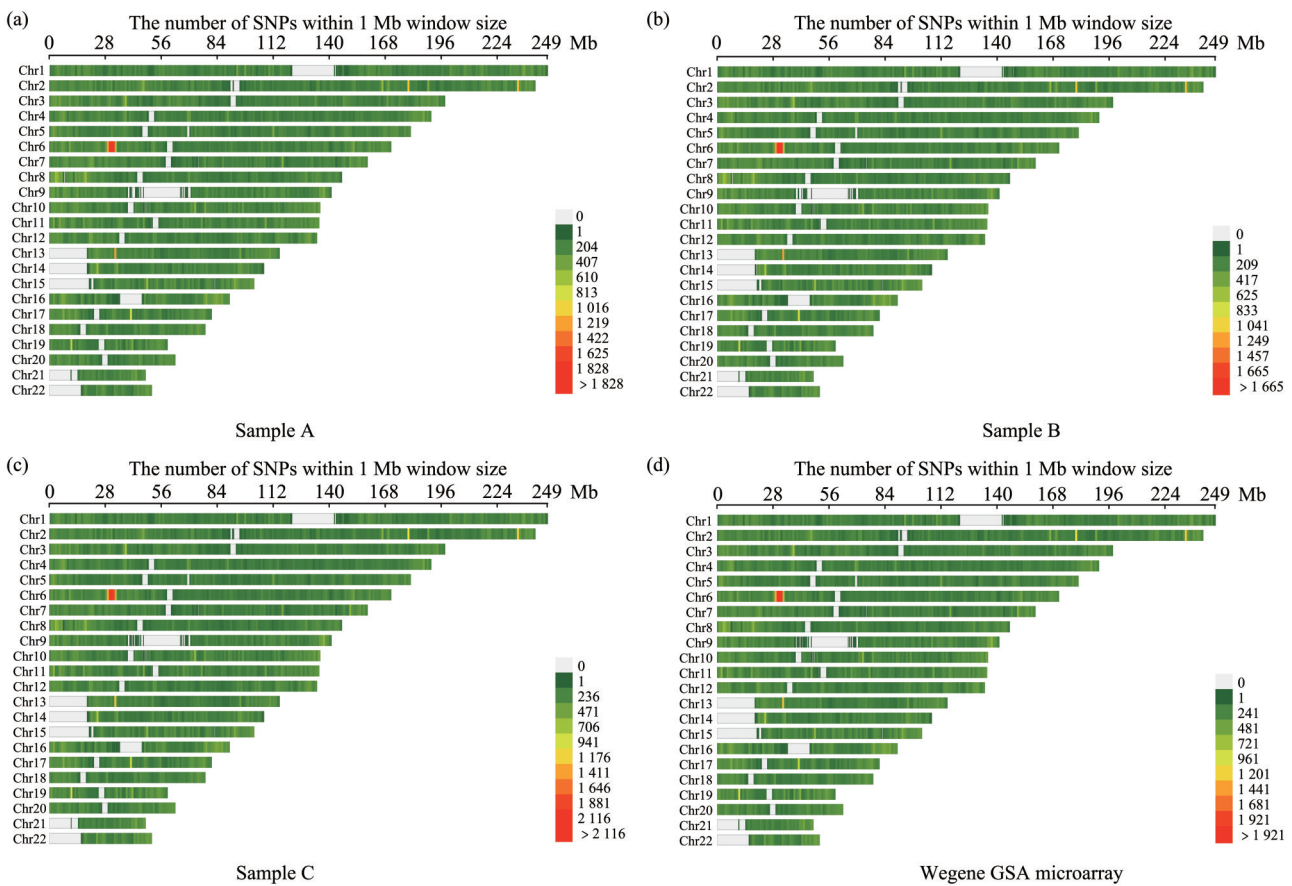


Fig. 5 Distribution density of SNP in sequenced samples on autosomes

表6中展示了合并后数据进行IBS亲缘关系预测的准确性, 本文把8级及8级以上亲缘定义为8th。从表中可以看出, 1~4级亲缘关系的CIA为100%。由于IBS算法的局限性, 随着亲缘关系等

Table 6 The evaluation of the accuracy of genetic relationship prediction by IBS

Real kinship	IBS Predict kinship									AC/% ¹⁾	CIA/% ²⁾	FN/% ³⁾	FP/% ⁴⁾
	PO/FS	2nd	3rd	4th	5th	6th	7th	8th	UN				
PO/FS	9									100	100	0	
2nd		14								100	100	0	
3rd			1	6						85.7	100	0	
4th				1	14	4				73.7	100	0	
5th						15	3	2	1	68.2	81.8	4.5	
6th							3	6	4	20	43.3	43.3	
7th									1	0	0	90.9	
8th												100	
UN									3				2.8

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 8th 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.

级的增加, 预测准确率也随之降低, 5级置信区间准确率降为81.8%。5级开始出现假阴性 (FN), 5级之后的亲缘关系绝对准确率 (AC) 明显下降。假阳性率 (FP) 为2.8%, 将无关亲缘预测为8级亲缘。30×测序数据使用IBS算法能够预测四级及以内的亲缘关系。

表7中展示了合并后的数据进行IBD亲缘关系预测的准确性。表中显示, IBD算法预测1~5级亲

缘关系的置信区间准确率为100%, 仅在预测8级及以上关系时出现假阴性, 4级以上绝对准确率开始下降。假阳性率为2.8%, 将无关亲缘预测为8级亲缘。因此, 30×测序数据使用IBD算法能够用于7级及以内亲缘关系预测, 将7级及以下亲缘预测为有关系亲缘的准确率为100%。图6显示, 30×测序数据使用IBD算法获得亲缘关系对的共享IBD长度与芯片数据基本一致。

Table 7 The evaluation of the accuracy of genetic relationship prediction by IBD

Real kinship	IBD predict kinship									AC/%	CIA/%	FN/%	FP/%
	PO/FS	2nd	3rd	4th	5th	6th	7th	8th	UN				
PO/FS	9									100	100	0	
2nd		14								100	100	0	
3rd			7							100	100	0	
4th			1	14	4					73.7	100	0	
5th				2	18	2				81.8	100	0	
6th					3	19	6	2		63.3	93.3	0	
7th							2	9		18.2	18.2	0	
8th								6	9			60	
UN								3	104				2.8

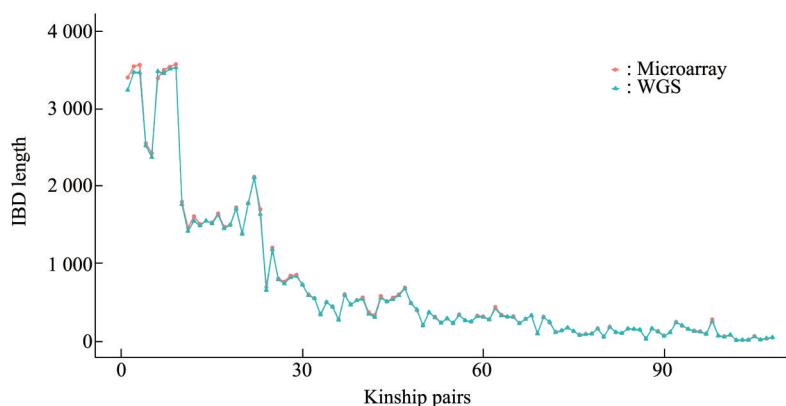


Fig. 6 Share IBD length distribution in kinship pairs

经 t 检验, 30×测序数据获得的IBD片段长度与芯片数据对比, P 值为0.93, 远大于0.05, 无显著差异。WGS数据可用于亲缘关系预测 (图7)。

2.5 族群推断

图8显示千人数据库5大人群共2 504个个体与4个质控测序样本合并数据的主成分分析结果, 左

下为欧洲人群, 左上为东亚人群, 标准品NA12878在参考数据中欧洲人群处聚集, 趋于左下方。中国北方样本A、B、C与东亚人群聚集, 都趋向于左上方。主成分分析结果与调查结果一致, 在全基因组层面能够区分出样本的族群来源。

从WGS数据中提取4份样本的27AISNP分型,

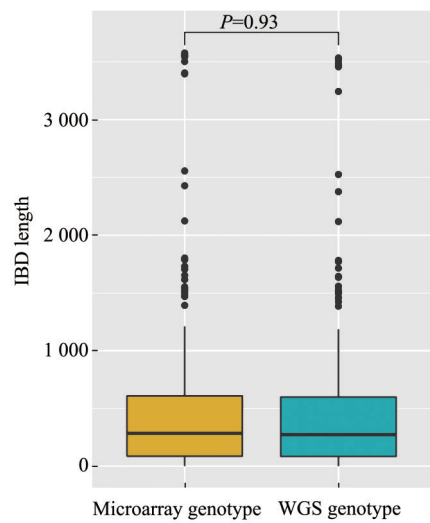


Fig. 7 t Test results

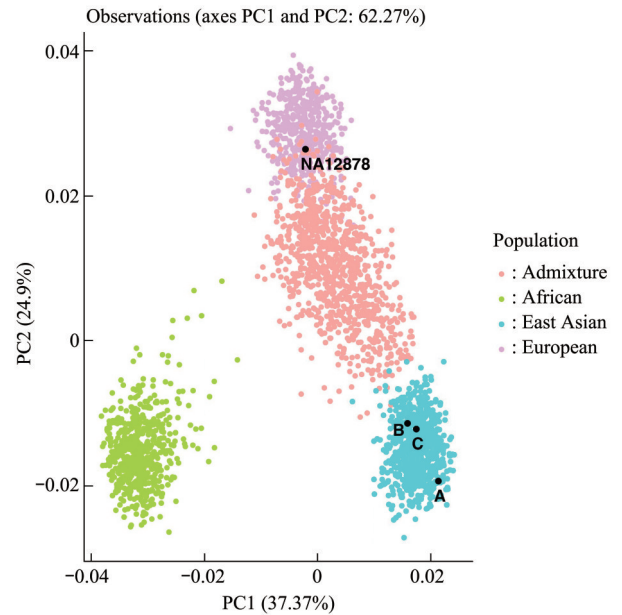


Fig. 9 27A1SNP inference ancestry results

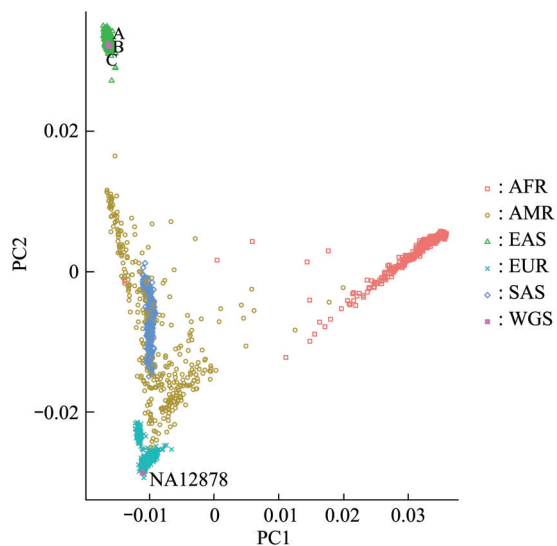


Fig. 8 Results of whole genome principal component analysis

4个样本的27SNP位点与一代测序分型完全一致, 准确率为100%。用该位点组合进行的族群推断结果如图9, 主成分1和主成分2共解释了总方差的62.27%, 与全基因组主成分分析结果一致。表明能够通过从WGS数据中提取特定位点集合的分型, 准确预测出族群来源。

3 讨 论

亲缘个体之间拥有相同的且来自共同祖先的基因片段称为IBD, 随着减数分裂次数增加, 非同源染色体重组次数增加, 所以亲缘关系越远, IBD越短。法医SNP系谱推断就是基于此原理进行亲缘关系的分析^[9]。高密度SNP数据是进行系谱推断的基础。在项目组前期研究中发现, 针对微量降解检材, SNP芯片的检出率较低^[10]。而WGS对法医中常见的微量降解检材具有优势, 能获得较多的位点分型。国外有研究利用ThruPLEX®试剂盒针对仅50 pg片段化DNA构建文库, 进而WGS获得SNP进行系谱推断并破获案件^[18]。针对高度降解的毛干等检材, WGS也能检测出大量SNP分型数据用于法医SNP系谱推断^[19]。目前, 国内尚无WGS技术用于法医系谱案件的研究报道。因此, 本研究在国产MGISEQ-200RS平台上构建WGS法医SNP系谱推断体系, 并对其位点分型准确性和预测准确性进行探讨。

本研究通过标准品的测序数据与千人基因组数据对比, A、B、C与芯片数据对比, 探讨了高精度WGS的一致性。以深度为4质量值为10质控过

滤,在保证检出率的条件下明显降低假阴性和纯合不一致位点数量,标准品质控后检出率为99.21%,一致率达99.84%,不一致率低于Tillmar等^[18]预测的30×深度不一致率0.8%。提取率高于Bode Technologies研讨会“Forensic Genealogy: Unlocking the Science of Genealogy”所研究的,血液WGS起始量250~50 ng提取率在97%~87%之间的范围,较GSA芯片的检出率100%~95%低,但对极低起始量(1~0.25 ng)样本,WGS提取率的稳定性远高于GSA芯片^[9]。对来自3个不同家系的3份样本进行平均深度30×的WGS,与芯片分型数据对比,SNP分型一致性达99.62%以上。有研究表明,高深度WGS数据中假阴性位点集中在低深度位点部分,且低频突变($0.5\% < \text{MAF} < 5\%$)和罕见突变($\text{MAF} < 0.5\%$)的一致性率较 $\text{MAF} > 5\%$ 的位点低^[20],解释了在设置质控标准后假阴性位点减少的现象。值得注意的是,质控标准过高可能导致一致性并没有显著提升,反而检出率过低影响后续分析,因此需要依据数据的应用方向进行评估。质控过滤后的SNP位点与芯片的SNP位点分布一致,WGS不存在位点分型的特异性问题,在全基因组均匀分布适用于法医高密度SNP系谱推断。

目前,已有针对全基因组SNP芯片数据预测亲缘关系准确性的系统研究。起始量200 ng的Illumina GSA芯片分型数据IBD算法预测7级亲缘,准确性为74.6%^[7]。Wgene GSA芯片分型数据IBS算法预测4级亲缘,绝对准确性为68.0%,置信区间准确率为98.1%^[6]。在2020年,国外开始出现WGS获取积案中骨骼SNP数据用于系谱推断的报道,获得平均深度32.3×共1 035 274个SNP,在第三方数据库中得到IBD大于10 cM的36个潜在亲缘,但未能调查证实潜在亲缘的真实性^[18];WGS检测陈旧血迹获得60×数据共1 861 000 SNP分型,在第三方数据库检索到共享IBD为347 cM的潜在亲缘,通过侦查锁定嫌疑人,利用STR图谱认定为作案人^[21]。但除了上述WGS系谱推断的个案报道,国内外尚无在真实家系中研究不同等级亲缘推断准确性的系统研究。由于系谱推断的结果将成为指导侦查方向的重要依据,为避免造成方向错误和浪费警力,非常有必要针对该技术进行系统研究,评估预测结果准确性。通过本研究发现,使用IBS算法,可用于1~4级亲缘关系的预测,4级置信区间准确率达100%。4级亲缘推断准确性较已有研

究^[6]稍高且无假阴性,可能是由于4级关系对数量较少,预测不准确的小概率事件并没有发生。使用IBD算法,能够对1~7级亲缘进行预测,7级及以内亲缘预测为有亲缘的准确率达100%,假阴性为0%。IBD算法对预测8级亲缘也具有一定的价值,即使在IBD片段长度不能达到预测分级的标准时,也可作为案件侦查的一个排查方向。经 t 检验,30×测序结果与芯片预测结果无显著差异,可用于法医系谱推断。

本研究还探究了使用WGS用于族群地域推断的可行性,全基因组SNP位点和提取AISNP体系位点推断结果均与调查结果一致,即WGS法医SNP系谱推断体系可用于DNA供者的族群来源推断。

4 结 论

基于国产测序平台的WGS获取的SNP位点组合与全基因组SNP芯片的分型一致率高,可与全基因组SNP芯片数据合并分析,进行准确的系谱推断。此外,本研究使用的测序仪器、耗材和试剂均为国产,数据安全性高,易于在法医实验室普及。下一步将继续优化体系性能,探索并建立不同测序深度的测序技术对微量降解检材的检测能力,以及不同等级亲缘关系的预测性能。

参 考 文 献

- [1] 李彩霞,胡兰.法医遗传学研究新进展.中国法医学杂志,2013,28(2):126-129
Li C X, Hu L. Chin J Forensic Med, 2013, 28(2): 126-129
- [2] 陈慧,李燃,臧钰,等.祖孙关系鉴定中引入不同参考个体的系统效能分析.中国法医学杂志,2022,37(1):24-29
Chen H, Li R, Zang Y, et al. Chin J Forensic Med, 2022, 37(1): 24-29
- [3] Phillips C. The Golden State Killer investigation and the nascent field of forensic genealogy. Forensic Sci Int: Genet, 2018, 36: 186-188
- [4] 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
- [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 Sci Technol, 2021, 46(6): 652-656
- [6] 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
Guan S S, Zhang W J, Wei Y L, et al. Prog Biochem Biophys, 2022, 49(3): 591-599

- [7] 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
- [8] DBM, Odile L, A I J. Fragmented nuclear DNA is the predominant genetic material in human hair shafts. *Genes*, 2018, **9**(12): 640
- [9] Kling D, Phillips C, Kennett D, *et al.* Investigative genetic genealogy: current methods, knowledge and practice. *Forensic Sci Int Genet*, 2021, **52**: 102474
- [10] 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
- [11] 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
Liu J, Ji A Q, Li C X, *et al.* *Forensic Sci Technol*, 2019, **44**(3): 189-194
- [12] Greytak E M, Moore C, Armentrout S L. Genetic genealogy for cold case and active investigations. *Forensic Sci Int*, 2019, **299**: 103-113
- [13] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 2015, **526**(7571): 68
- [14] 魏丽, 魏以梁, 刘京, 等. 27 个常染色体 AIM-SNP 的筛选和验证. *中国法医学杂志*, 2017, **32**(3): 235-239
Wei L, Wei Y L, Liu J, *et al.* *Chin J Forensic Med*, 2017, **32**(3): 235-239
- [15] Guangbin Z, Guanju M, Chi Z, *et al.* BGISEQ-500RS sequencing of a 448-plex SNP panel for forensic individual identification and kinship analysis. *Forensic Sci Int Genet*, 2021, **55**: 102580
- [16] Kongyang Z, Panxin D, Jianxue X, *et al.* Comparative performance of the MGISEQ-2000 and Illumina X-Ten sequencing platforms for paleogenomics. *Front Genet*, 2021, **12**: 745508
- [17] Chen Y C. Effects of GC bias in Next-Generation-Sequencing data on *de novo* genome assembly. *PLoS One*, 2013, **8**(4): e62856
- [18] 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
- [19] Ida M, Sand K T, Andaine S, *et al.* Identifying a living great-grandson of the Lakota Sioux leader Tatanka Iyotake (Sitting Bull). *Sci Adv*, 2021, **7**(44): eabh2013
- [20] Hwang K B, Lee I H, Li H, *et al.* Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. *Sci Rep*, 2019, **9**(1): 3219
- [21] Andreas T, Aili F S, Jan S, *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

Study of The SNP Investigative Genetic Genealogy Based on Whole Genome Sequencing*

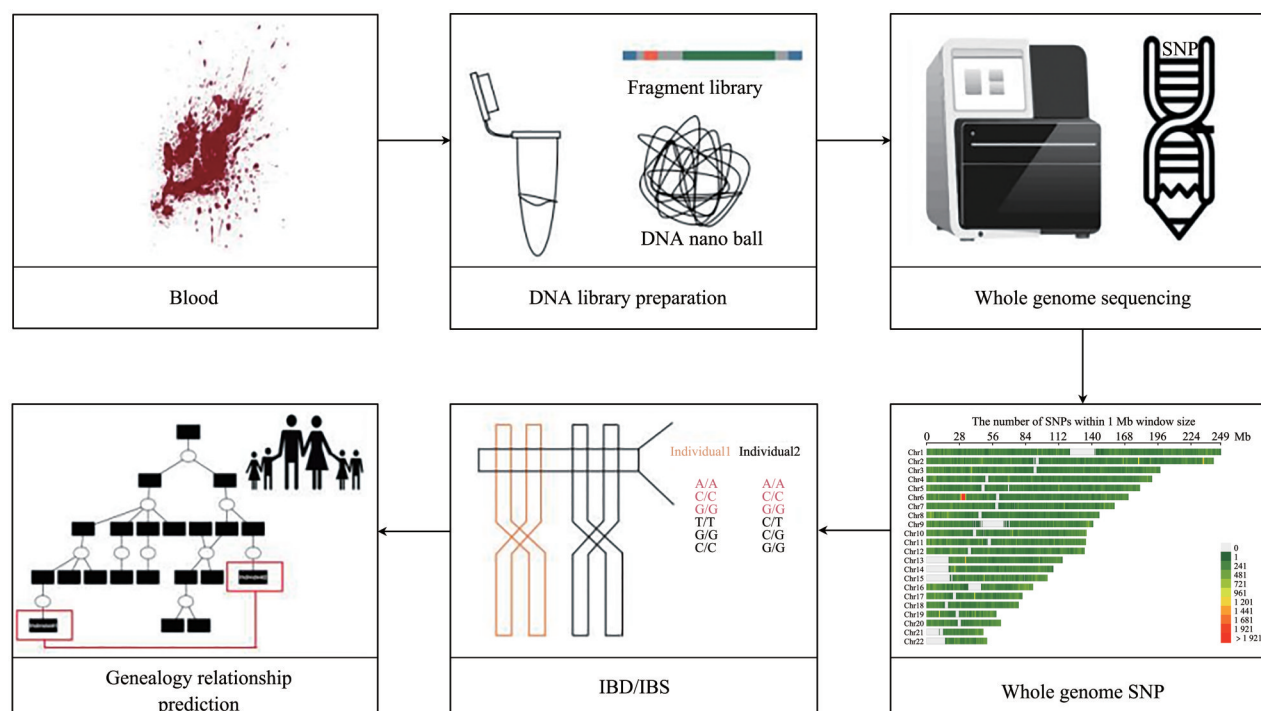
XIE Quan^{1,2)}, GUI Xu^{1,2,3)}, ZHAO Wen-Ting²⁾, FANG Zhi-Xiao²⁾, XU Jing-Yi²⁾,
LIU Jing²⁾, LI Cai-Xia^{1,2)}**

⁽¹⁾School of Investigation, People's Public Security University of China, Beijing 100038, China;

²⁾Beijing Engineering Research Center of Crime Scene Evidence Examination, National Engineering Laboratory for Forensic Science, Institute of Forensic Science, Beijing 100038, China;

³⁾Public Security Bureau of Dachuan District of Dazhou City, Dazhou 635000, China)

Graphical abstract



Abstract Objective The high density SNP genotype data were obtained by whole genome sequencing (WGS), the accuracy of genotype was evaluated, and this study is intended to establish the method of using whole genome sequencing data for SNP kinship relationship prediction. **Methods** The samples were sequenced at a depth of 30× through the MGISEQ-200RS sequencing platform, and 645 199 autosomal SNP in the Wegene GSA DNA

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

** Corresponding author.

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

Received: October 26, 2022 Accepted: March 1, 2023

microarray were extracted from the sequencing data. After quality control and filtering, the prediction relationship was calculated and predicted by IBS/IBD algorithm, and the ancestry of the samples was analyzed. **Results** The coincidence rate between SNP genotype extracted from sequencing data and Wegene GSA genotype is more than 99.62%. The SNP genotype obtained by sequencing can predict kinship from level 1 to level 4 by using IBS algorithm, and confidence interval accuracy of level 4 kinship prediction is 100%. Using IBD algorithm, the confidence interval accuracy of level 1 to 7 kinship prediction is 100%. The pedigree inference ability of SNP obtained from high-depth whole genome sequence data is not significantly different from that of DNA microarray prediction. At the same time, the use of whole genome sequencing data for ancestry inference is consistent with the survey results. **Conclusion** The whole genome resequencing technique can be applied to SNP genealogy inference to provide clues for case detection.

Key words single nucleotide polymorphism, whole genome sequencing, investigative genetic genealogy, kinship

DOI: 10.16476/j.pibb.2022.0506