



高密度 Y-SNP 分型体系用于法医实践中 Y 单倍群推断的效能评估*

张德琴^{1,2)} 王春年^{2,3)} 姜琳琳^{2,4)} 倪萌⁵⁾ 高静^{2,3)} 黄江^{1)**} 江丽^{2)**}

⁽¹⁾ 贵州医科大学毒性检测中心, 法医学院, 贵阳 561113;

²⁾ 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,
现场物证溯源技术国家工程实验室, 北京 100038;

³⁾ 中国人民公安大学侦查学院, 北京 100038; ⁴⁾ 山西医科大学法医学院, 晋中 030600;

⁵⁾ 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 Y染色体单倍群分配的准确性在男性样本的父系祖先溯源中至关重要。随着高通量测序技术的发展, 目前研究常基于全基因组或芯片数据中高密度Y染色体单核苷酸多态性(Y-SNP)的分型来判别个体的Y染色体单倍群。本研究系统评估了4种常见的高密度单核苷酸多态性(SNP)分型检测体系, 包括全球筛查芯片(Global Screening Array, GSA)、中国基因分型芯片(Chinese Genotyping Array, CGA)、Affymetrix芯片及1240K捕获体系的单倍群判别效能, 为不同体系的数据比对提供参考依据。**方法** 基于千人基因组计划1590例男性样本的30×全基因组测序(whole genome sequencing, WGS)数据提取了上述4种Y-SNP位点集的分型结果, 收集了109个中国汉族家系的384对1~12级亲缘样本的GSA芯片分型数据, 利用Y-LineageTracker v1.3.0软件进行单倍群判别。评估4种Y-SNP位点集与全基因组分型数据获取的单倍群一致性、分辨率。评估千人基因组计划及本研究收集的109份家系样本的单倍群一致性、分辨率。评估不同量级Y-SNP对单倍群判别结果的影响。**结果** GSA和CGA位点集在判别单倍群亚群和分辨率上优于其他两种位点集; GSA、CGA和1240K位点集的单倍群判定与全基因组分型数据一致性均超过88.70%, 而Affymetrix平台的一致性显著降低至61.89%。所有家系样本中4组位点集均未出现单倍群不匹配亲缘关系对。在1000~10000个Y-SNP范围内, Y-SNP位点组合数量与分类一致性呈正相关, 但当Y-SNP数量超过10000时, 其分类一致性趋于平缓。**结论** GSA和CGA位点集兼具高分辨率与高一一致性, 能够在法医实践中为Y单倍群判定提供可靠依据。

关键词 Y-SNP, NRY单倍群分类, 父系谱系, 祖先溯源

中图分类号 R89, D919.2, Q39

DOI: 10.3724/j.pibb.2025.0481

CSTR: 32369.14.pibb.20250481

人类Y染色体在研究人类进化和遗传学研究领域具有关键作用^[1-2]。目前最常用的Y染色体遗传标记有短串联重复序列(short tandem repeat, STR)和单核苷酸多态性(single nucleotide polymorphism, SNP)。相较于Y-STR, Y-SNP突变率极低, 约为 1.0×10^{-9} 次突变/代^[3]。Y染色体非重组DNA区域(non-recombining region, NRY)占据了Y染色体总长度的95%, 在遗传过程中, 该段DNA不与X染色体发生同源重组, 直接由父亲传递给儿子, 仅通过突变而改变, 确保了其遗传信息的稳定性和连续性^[4]。因此, Y染色体上的变异模式被广泛用于揭示人类迁移的历史轨迹, 在人类进

化起源、遗传学、法医学以及系谱重建等研究中具有重要意义^[5-8]。

到目前为止, 国际遗传系谱学会(International Society of Genetic Genealogy, ISOGG)在NRY中汇编了所有变异, 覆盖了20个主要单倍群(A~T), 每个主要单倍群下通过特定

* 国家重点研发计划(2022YFC3341004), 国家自然科学基金(81772027)和公安部科技计划(2023JSZ02)资助项目。

** 通讯联系人。

黄江 Tel: 0851-88596769, E-mail: mmm_hj@126.com

江丽 Tel: 010-83752707, E-mail: jl@mail.bnu.edu.cn

收稿日期: 2025-11-03, 接受日期: 2025-12-26

Y-SNP标记细分为多个亚单倍群。Y染色体上存在数千个能提供系统发育信号的SNP和小片段插入缺失(Indel), 可用于构建进化树并识别单倍群谱系。这些单倍群呈现地理分布与民族语言特异性, 使其在推断历史群体结构和父系祖先溯源方面具有重要价值^[9-10]。主干单倍群呈现明显全球分布差异, 其中A单倍群作为Y染色体系统发育体系中最古老的根基分支, 常分布于非洲地带, 而单倍群O在东亚占主导地位^[11-13]。R1a等特定支系呈多区域分布模式: 其一主要分布于欧洲大陆, 其二则局限于中亚及南亚地域, 暗示其可能起源于欧亚草原后逐渐迁移扩散。F单倍群则主要见于印度地区^[14-15]。亚洲各地广泛存在的C3单倍群, 往往被视为成吉思汗帝国血缘遗传的历史印记^[16]。Y单倍群在历史身份验证中价值显著: 对法国国王亨利四世与路易十六的三位父系后代分析显示, 其共享单倍群与据称属于这两位国王的遗骸不一致, 对其真实性提出质疑^[17-19]。此外, 通过比较父系谱系, 反驳了中国皇帝曹操(公元155年~公元220年)与曹参侯爵有血缘关系的说法^[20-21]。因此, Y单倍群能否被准确判别在远距离父系祖源溯源研究中是一个非常关键的问题。

高通量测序技术(high-throughput sequencing, HTS)自20世纪90年代末迅速发展, 凭借其短时间内生成海量基因组数据的能力, 成为现代基因组学的核心技术, 彻底改变了基因组学领域^[22-23]。目前主流的HTS技术为第二代测序(next generation sequencing, NGS), 其通量高、成本低、灵敏度强, 可同时检测数百至数十万个单核苷酸多态性位点^[24-25]。靶向富集与微阵列基因分型技术的进步进一步优化了成本效益。基于杂交探针捕获的靶向测序方法与SNP分型技术均能实现特定基因组区域的高效、高质量变异检测^[26-28]。其中全球筛查芯片(Global Screening Array, GSA; illumina)包含22 512个Y-SNP位点, 涵盖了全球26个人群常见变异, 中国基因分型芯片(Chinese Genotyping Array, CGA; illumina)针对东亚人群设计, 包含24 047个Y-SNP位点, Affymetrix(23MF_V2.2 High-Density SNP Array)芯片^[29](由23魔方公司根据Affymetrix公司生产的Precision Medicine Research Array(PMRA)芯片)修改后, 包含了27 300个Y-SNP位点, 古DNA领域多依赖The Allen Ancient DNA Resource(AADR)的1240K探针捕获体系进行人类学分析^[30], 包含

32 670个Y-SNP位点, 其侧重于祖先变异的分析。国内外研究人员还开发了Y-SNP检测体系以应用于生物地理祖先推断。例如: 荷兰Kayser团队^[31]于2019年开发的包含859个Y-SNP的检测体系已实现商业化; 2020年, Claerhout等^[32]报道了15 611个Y-SNP的检测体系, 但未公开完整位点信息; 一些国内学者团队则聚焦中国人群系统发育树, Wang等^[33-34]于2022~2023年构建了包含256个、639个Y-SNP的体系及Li等^[13]构建的381个Y-SNP检测体系。目前在商业化应用方面, 23魔方、23andMe等DTC公司提供定制Y-SNP检测或全基因组测序(whole genome sequencing, WGS), 结合YTree(<https://www.yfull.com/tree/>)为用户提供父系祖源溯源分析。然而, 基于这些不同面板(panel)数据得到的单倍群进行父系祖源溯源的一致性仍然有待进一步评估, 为不同检测平台的数据比对带来了困难。因此, 本文基于千人基因组30×高深度测序项目中来自全球26个人群的1 590份男性样本和本实验室109份汉族男性样本的GSA芯片分型数据, 从单倍群分类一致性、分辨率大小等多个方面分析比较GSA位点集、CGA位点集、Affymetrix位点集以及1240K位点集的Y染色体单倍群划分效能和人群区分能力, 此外, 还评估GSA芯片分型数据在远近亲缘家系样本中分类一致性以及不同量级Y-SNP位点组合的Y单倍群分类一致性, 以期跨平台的Y单倍群数据比对提供依据, 为父系祖源溯源和系谱研究提供参考。

1 材料与方法

1.1 数据来源

本研究收集了3组数据: 数据集1为千人基因组1 590人份的男性30×WGS样本^[24]; 数据集2为千人基因组609份1级亲缘样本, 数据集1和数据集2人群信息如表S1所示; 数据集3包含本实验室109例通过WeGene GSA v2.0芯片分型的中国汉族男性样本^[35], 所有参与者均签署了知情同意书, 并通过了公安部鉴定中心伦理委员会批准(编号: 2023-004)。

1.2 数据处理

使用VCFtools v0.1.17软件^[36]对数据集1、数据集2进行了样本过滤, 过滤条件设置: 样本基因型缺失率小于0.1, 位点分型缺失率小于0.1。最终数据集1和数据集2包含181 601个Y-SNP。4组位点集分别为WeGene GSA v2.0(以下简称为GSA)、

WeGene CGA v3.0 (以下简称为 CGA)、23MF_V 2.2 High-Density SNP Array (以下简称为 Affymetrix) 和 The Allen Ancient DNA Resource (AADR) 中的 1240K Version v54.1.p1 (以下简称为 1240K)。数据集 1、数据集 2 为 GRCh38 版本, GSA、CGA、1240K 和 Affymetrix 位点集为 GRCh37 版本, 为确保参考基因组坐标一致, 使用 LiftOver 软件^[37] (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>) 将 4 组位点集基因组坐标从 GRCh37 转换为 GRCh38 版本。利用 bcftools v1.10.2 软件^[37]

从数据集 1 和数据集 2 中提取上述 4 组位点集的 Y-SNP 分型数据, Y 单倍群的判别基于突变的类型, 包括 SNP 或者 Indel, 因此若该位点在所有样本中均未发生突变, 提取 Y-SNP 分型时该位点则被排除。为比较 4 组位点集提取到的 Y-SNP 与 ISOGG 谱系树交叉位点占比及特异性位点, 使用 Rstudio v4.4.0 中的 ggVennDiagram 包绘制了维恩图 (图 1), 并使用 CMplot 包绘制 SNP 密度图展示 4 组位点集 Y-SNP 在 Y 染色体分布情况 (图 2)。

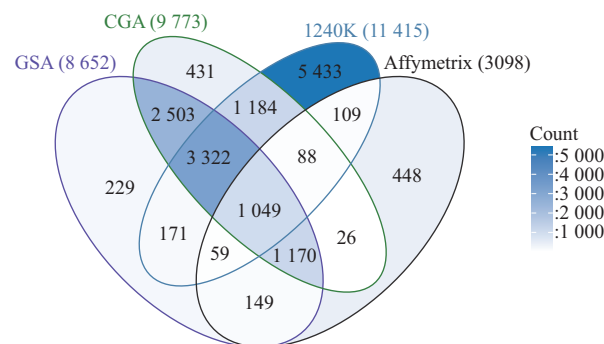


Fig. 1 Venn diagram illustrating the overlap between 4 Y-SNP panels extracted from WGS and ISOGG database

After interconversion of genomic coordinates, the respective sizes of the GSA, CGA, 1240K, and Affymetrix marker sets are 22 512, 24 047, 32 658, and 27 299 loci, all classified into ancestral and derived types. The inference of Y-haplogroups is independent of ancestral loci, being determined solely by the genotyping of derived loci. In the 1 590 samples, the number of invariant ancestral loci for each set was 12 346, 12 179, 17 012, and 15 334, respectively.

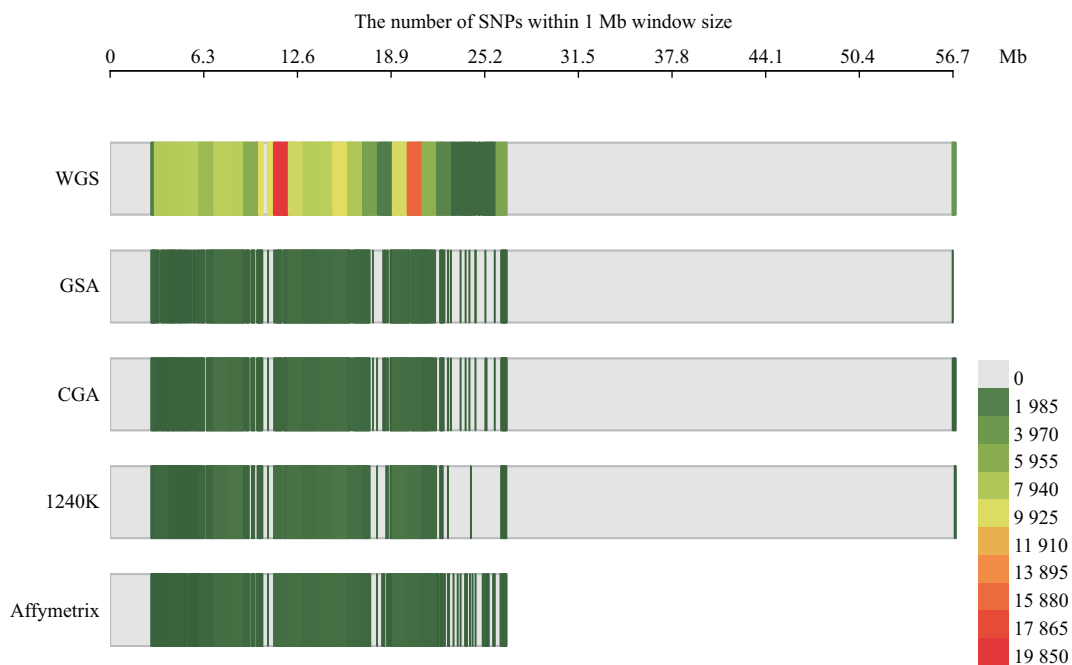


Fig. 2 Distribution map of each Y-SNP on Y chromosomes of 4 panels and WGS

1.3 单倍群分类一致性比较

为评估各Y-SNP位点集的单倍群分类一致性及分配结果的准确可靠性, 将具有更优测序深度和覆盖度的数据集1设为参考标准。通过Y-LineageTracker v1.3.0软件^[38]基于Y染色体谱系树(ISOOG网站版本号15.73)进行单倍群分配, 并采用直接计数法统计单倍群分布。设定以下指标量化单倍群分类一致性(其中 N 为样本总量, X 为4组位点集单倍群分配结果与数据集1完全匹配的样本数, Y 为单倍群分配结果与数据集1存在谱系上下游关系的样本数):

a. 完全一致率 (complete consistency rate, CCR)。CCR= X/N , 即为, 不同位点集数据与数据集1判别的单倍群完全匹配的样本比值。

b. 不完全一致率 (incomplete consistency rate, ICR)。ICR= Y/N , 即为, 不同位点集数据与数据集1判别单倍群存在上下游谱系关系的样本比例。

c. 错判率 (misclassification rate, MR)。MR= $(1 - X - Y) / N$, 即为, 不同位点集与数据集1单倍群分配结果非同一直干单倍群或非上下游谱系关系的样本比例。

d. 一致性 (consistency, C)。C=CCR+ICR, 完全一致率与不完全一致率之和。

使用Rstudio v4.4.0中的ggplot2软件包绘制不同位点集判别单倍群结果的洲际分面柱状图。为评估4组位点集在不同单倍群中的判别效能差异, 基于WGS数据判别单倍群, 将样本量低于50的单倍群排除, 统计不同单倍群(如O-M175、R-M207和Q-M242单倍群)下4组位点集的CCR、ICR以及MR, 使用Rstudio v4.4.0中的ggplot2软件包绘制4组位点集在不同单倍群中的分面柱状图。

家系样本使用数据集2提取的GSA位点集、CGA位点集、Affymetrix位点集和1240K位点集Y-SNP以及数据集3, 按上述公式计算错判率和一致性评估不同远近父系亲缘等级之间单倍群一致性。

1.4 分辨率评估

为了实现不同位点集的单倍群划分精细程度的比较, 本研究引入了“单倍群分辨率”作为评估指标。一个字母或数字代表单倍群分辨率为1级。主单倍群(A~T)下亚群分支长度直接反映了单倍群划分的精细程度, 即亚单倍群分支越长, 系统发育树上的位置也越接近末端^[39]。例如相比于单倍群R1a(3级), R1a2a1b(7级)属下游单倍群则分辨

率更高。本研究对数据集1中所有样本进行分析, 使用直接计数法统计WGS与4组位点集判别的末端单倍群类型即分辨率等级。基于SPSS 27.0 (IBM, USA)统计软件, 采用非参数检验中曼-惠特尼U检验(Mann-Whitney U test)方法对不同Y-SNP位点集与全基因组数据集的分辨率差异进行显著性检验并使用Rstudio v4.4.0中的ggplot2软件包绘制箱线图评估分辨率大小。

1.5 不同量级Y-SNP位点组合评估

从数据集1与ISOOG谱系树重叠位点中对1590份样本按照设定的Y-SNP位点数量用python脚本进行随机抽样, 分别抽取1000、5000、10000、15000、20000、25000个Y-SNP, 每组位点抽样流程重复100次, 并用Plink提取对应Y-SNP的分型数据。观察Y单倍群准确性与Y-SNP数量之间的关系; 基于上述结果, 对GSA、CGA位点集与ISOOG谱系树的交叉位点进行500、1000、2000、3000、4000、5000、6000、7000、8000个Y-SNP的随机抽样, 每组位点抽样流程重复100次。以数据集1作为参照基准, 针对不同量级Y-SNP位点组合评估单倍群准确性(参数同1.3)并绘制柱状折线混合图和带误差线的柱状图。

2 结 果

2.1 不同位点集分类一致性

对数据集1进行了4组位点集的提取与单倍群推断, 并与数据集1所分类的单倍群进行了比较。基于1590份样本, GSA、CGA、1240K、Affymetrix位点集所判别的单倍群与基于全基因组判别的单倍群一致性分别为88.74%、92.08%、92.96%、61.89%, 其中GSA、CGA、1240K位点集单倍群一致性均达到了88.00%以上。GSA、CGA、1240K、Affymetrix位点集完全一致率(单倍群完全匹配)分别为63.90%、64.78%、26.35%、6.54%。尽管1240K位点集一致性较高, 但其完全一致率相比GSA位点集、CGA位点集显著降低。4组位点集的洲际比较分面柱状图(图3)显示, GSA位点集、CGA位点集以及1240K位点集的整体错判率较低, 其中, GSA位点集和CGA位点集错判单倍群主要分布在非洲人群, 1240K位点集错判单倍群则主要为美洲人群。Affymetrix位点集错判率则较高, 主要分布在东亚人群。值得注意的是, CGA位点集和GSA位点集在东亚地区错判较少, 且CGA位点集在中国人群中所判别的单倍群

一致性达 100% (表 S2), 这与 CGA 位点集在中国人群中的特异性密切相关。

基于单倍群 O-M175 和 H-L901, GSA 和 CGA 位点集的一致性优于其他两种位点集, 对这两个单倍群的判别一致性达到 100%。相比之下, Affymetrix 位点集将所有 O-M175 单倍群的个体错误归类到 K2-M526 单倍群, 而对 H-L901 单倍群的判定一致性仅为 1.41%, 其中 1.41% 的 H-L901 单倍

群被错判为 J-M304, 另外 97.18% 的 H-L901 单倍群被错判为 F-M89。基于 R-M207 单倍群, 4 组位点集一致性均超过 94.00%。值得注意的是, 对于 E-M96 单倍群, 1240K 和 Affymetrix 位点集一致性首次优于 GSA 和 CGA 位点集。基于 J-M304、Q-M242 和 I-M170 单倍群, 4 种位点集均存在相对较高的错判率, 其中 Affymetrix 位点集的整体错判率最高 (图 S1)。

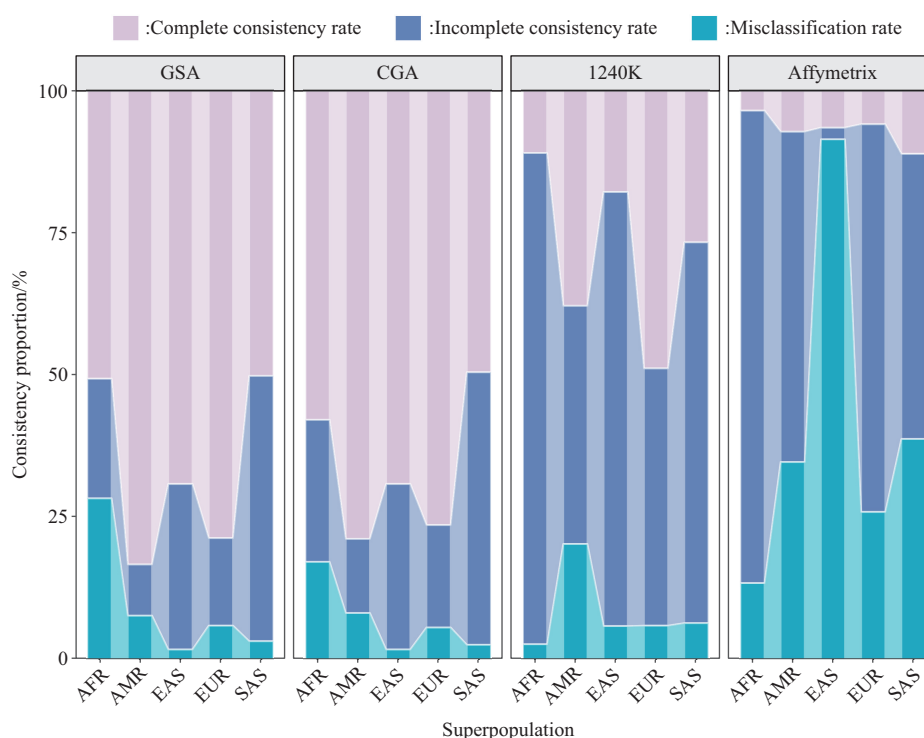


Fig. 3 Intercontinental comparison of different Y-SNP panels displayed as a faceted bar plot, based on Dataset 1

2.2 家系样本分类一致性

全基因组数据中包含 306 对 1 级亲缘关系的样本, 其中有 305 对父子与 1 对兄弟。亲缘关系样本单倍群比较结果显示。GSA 位点集、CGA 位点集、1240K 位点集和 Affymetrix 位点集对一级亲缘关系样本单倍群分类一致性达 100% (表 1)。数据集 2 存在 1 对单倍群不一致的亲缘样本, 其中父亲 Y 单倍群判别为 E1b1a1a1a1c2c3a3b, 儿子判别为 E1b1a1a1a1c2c3a3a。尽管其单倍群不一致, 但仅在单倍群末端分支存在 1 级差异。

数据集 3 包含 384 对 1 至 12 级亲缘关系样本, 单倍群分类结果显示, 1~12 级亲缘样本单倍群分类一致性均为 100% (表 2), 表明 GSA 芯片分型数

Table 1 Consistency of haplogroups between Dataset 2 and various Y-SNP panels in first-degree relative samples

Dataset	Kinship pairs	Concordant pairs	Misclassification rate/%	Consistency/%
WGS	306	305	0.33	99.67
GSA	306	306	0.00	100.00
CGA	306	306	0.00	100.00
1240K	306	306	0.00	100.00
Affymetrix	306	306	0.00	100.00

据对远至 12 级的亲缘关系样本仍然能够有效判别出单倍群。

2.3 分辨率评估结果

基于数据集 1, WGS 识别出 16 个主干单倍群

Table 2 Consistency of haplogroups from Dataset 3

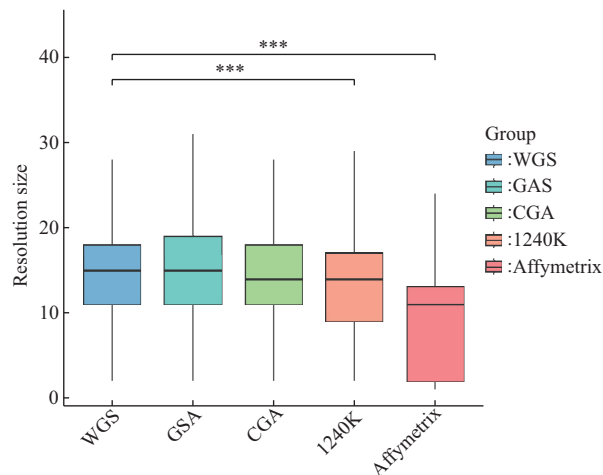
Kinship degree	Kinship pairs	Concordant pairs	Misclassification rate/%	Consistency/%
1	82	82	0.00	100
2	33	33	0.00	100
3	34	34	0.00	100
4	44	44	0.00	100
5	39	39	0.00	100
6	36	36	0.00	100
7	25	25	0.00	100
8	9	9	0.00	100
9	10	10	0.00	100
10	29	29	0.00	100
11	30	30	0.00	100
12	13	13	0.00	100

下 984 种末端单倍群, GSA、CGA、1240K 和 Affymetrix 位点集分别识别出 15 个主干单倍群下 778、803、506 和 158 个末端单倍群 (表 S3)。基于单倍群分辨率的箱线图结果 (图 4) 显示, 数据集 1 中 GSA 位点集所分类的单倍群分辨率效果最佳, 其次是 CGA 位点集所分类的单倍群, 分辨率较低的是 1240K 位点集与 Affymetrix 位点集所分类的单倍群。曼-惠特尼 U 检验显示 GSA 和 CGA 位点集与 WGS 的分辨率无统计学差异 ($P>0.05$), 而 1240K 与 Affymetrix 位点集与 WGS 的分辨率差异具有显著性 ($P<0.05$)。GSA 和 CGA 位点集对样本末端单倍群识别能力和单倍群分辨率高于另外 2 组位点集。

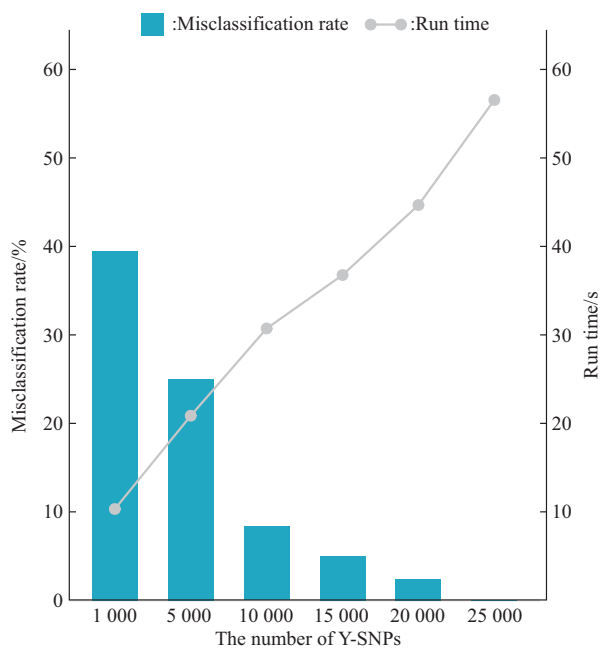
2.4 不同数量位点组合评估结果

在对数据集 1 随机提取的 Y-SNP 数据进行单倍群分析时, 观察到在 Y-SNP 数量介于 1 000~10 000 时, 随着位点数增加, 单倍群分类结果一致性也随之增加, 当 Y-SNP 数量超过 10 000 个时, 即使数量进一步成倍增加, 单倍群一致性的提升变得相对缓慢, 错判率也趋于稳定 (图 5)。此外, 分析的运行时间随着 Y-SNP 数量的增加而呈现出线性增加, 但在分析最多 25 000 个 Y-SNP 时所用时长仍在 60 s 以内。

GSA 和 CGA 位点集分别与 Y 染色体系统发育树 (ISOGG 15.73 版) 存在 8 652 和 9 773 个 Y-SNP 的交集位点。当 Y-SNP 数量在 500~1 000 时, 随着位点数量增加, 单倍群错判率变化甚微; 当 Y-SNP 数量在 1 000~5 000 个标记时, 随着位点数量增加,

**Fig. 4 Boxplots depicting haplogroup resolution across different Y-SNP panels from Dataset 1**

The panels contain the following Y-SNP counts: WGS (181 601), GSA (10 166), CGA (11 868), 1240K capture panel (15 646), and Affymetrix (11 965). Asterisks indicate statistical significance from ANOVA: *** $P<0.001$.

**Fig. 5 Combined bar and line chart showing misclassification rate and computation time for varying sizes of Y-SNP panel combinations based on Dataset 1**

单倍群错判率急剧下降; 而在 5 000~8 000 个 Y-SNPs 范围内, 随着位点数量增加, 错判率下降速度减缓 (图 6)。

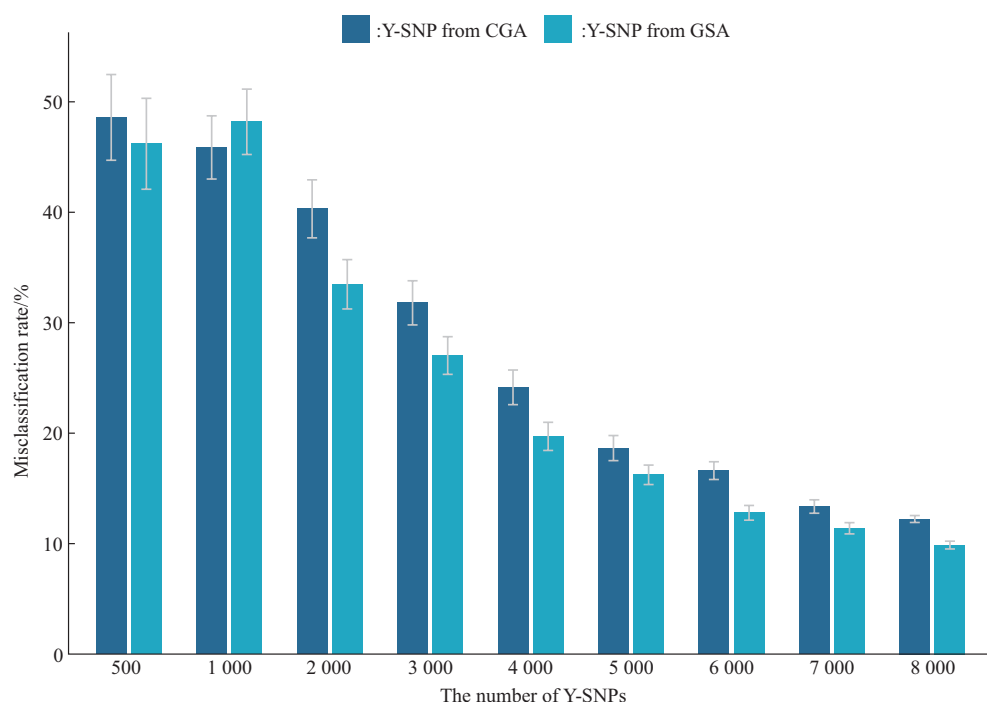


Fig. 6 Error-bar plot reveals misclassification rates of Y-SNP combinations randomly selected from loci shared among GSA, CGA panels and ISOGG database

3 讨 论

Y单倍群在全球范围内的分布格局能够有效反映人类历史进程中的迁徙轨迹与演化脉络^[40-41]。靶向富集技术和基因芯片技术的发展为人类提供了大量的基因组数据,基于Y染色体进行族群地域来源推断和父系祖先溯源的研究是研究热点之一^[42]。本研究整合了无关个体样本和家系样本数据,采用4种位点集(GSA、CGA、1240K和Affymetrix)进行单倍群推断与比较,并通过不同量级的Y-SNP位点组合验证单倍群判别的准确性。研究结果显示,在单倍群分类一致性和分辨率方面,GSA和CGA位点集均表现出良好的推断效能,相比之下,1240K位点集推断一致性较高但分辨率不足,Affymetrix位点集的推断效能则显著降低。在家系样本分析中,GSA、CGA、1240K和Affymetrix等4组位点集对一级亲缘样本单倍群分类一致性达100%,GSA芯片分型数据在不同远近亲缘关系等级(1~12级)的中国汉族样本中展现出较高的分类一致性,能够准确将样本归类至同一单倍群。进一步评估表明,单倍群分类一致性与Y-SNP数量呈正相关关系(1000~10000个Y-SNPs),超过10000个Y-SNPs时一致性趋于稳定。本研究为跨

平台的Y-SNP单倍群分析比对提供了重要参考依据。

3.1 不同位点集单倍群分类效能差异分析比较

NRV单倍群由特定的Y-SNP组合界定^[43],尽管少数研究利用Y-STR进行单倍群预测^[44-45],但由于该方法尚未成熟,目前主流研究仍主要依赖Y-SNP进行单倍群划分。不同位点集分类效能比较中,GSA、CGA和1240K位点集整体错判率较低,主要分布在非洲人群和美洲人群,而Affymetrix位点集的错判率最高,主要集中在东亚群体。GSA芯片为大规模群体遗传学研究设计,位点选择覆盖欧洲、东亚等全球多样化人群,有效避免了族群偏倚,常用于临床筛查易致病的常见基因型^[46-47]。CGA芯片则进一步优化了位点使其更适用于中国群体^[48]。与全基因组数据判别单倍群的分析比较中,单倍群不一致样本在GSA、CGA、1240K位点集中下游分支出现差异而与WGS同属一个上游单倍群,例如E1b1b1a1a1c2和E1b1b1a1a2a1b,判别错误的单倍群主要为E-M96、J-M304、Q-M242和I-M170,而在单倍群R-M207、O-M175和H-L901具有较高的一致性。在Affymetrix位点集中,除单倍群E-M96和R-M207外,其他几组单倍群错判率均处于较高水平,这可能与Affymetrix位

点集和 ISOGG 谱系树仅有 3 098 个交叉 Y-SNP 有关。全基因组数据通常包含野生型(分型与参考基因组一致)和突变型(SNP、Indel), 其中单倍群的分化与突变型紧密相关。GSA、CGA、1240K 和 Affymetrix 位点集从全基因组数据分别提取到 10 166、11 868、15 646 和 11 965 个 Y-SNP, 但其中仅有一部分 Y-SNP 与 ISOGG 谱系树相交, 相交位点越少的位点集通常不能提供足够的变异信息以识别主干单倍群和下游分支, 从而导致单倍群判别错误。在法医学研究中, 精细 Y 单倍群在解析不同人群的遗传分化及其迁徙模式方面具有重要应用前景^[49]。Rootsi 等^[50]根据 R1a-M582 谱系发现, Ashkenazi Levites 并非起源于东欧, 其祖先谱系的地理来源应位于近东地区, 且该谱系很可能在流散时期之前的希伯来人群中就已存在。本研究中, WGS 在 1 590 份样本中能识别出 984 种末端单倍群, GSA、CGA、1240K 和 Affymetrix 位点集分别能判别出 778、803、506 和 158 种末端单倍群。GSA 位点集和 CGA 位点集的分辨率最接近 WGS 水平, 无显著统计学差异, 单倍群分辨率均值均为 15, 即可识别亚群至第 15 级的亚单倍群如 O2a1b1a2a1a1a1b, 适用于精细单倍群分析, 如计算不同亚群间的遗传距离、估算群体分化时间以及揭示父系遗传的复杂性与多样性^[51]。通过分析单倍群一致性和分辨率等指标, GSA 芯片和 CGA 芯片虽然仅覆盖部分基因组区域, 但仍能提供足够的遗传信息用于家谱分析和祖源推断。这些发现为未来 Y 染色体遗传学研究中的位点集选择提供了重要参考。此外, 本研究涉及的相关 Y-SNP 位点、单倍群与潜在的临床疾病关联仍有待深入探索, 后续研究将利用肿瘤基因组药物系统(tumor genomic drug database, TGDD)等数据库^[52]对相关位点进行深度注释与分析, 以探索其在疾病风险评估中的潜在价值。

3.2 家系样本单倍群分类效能评估

基于数据集 2 提取的 4 种 Y-SNP 分型面板(GSA、CGA、1240K 和 Affymetrix)在家系样本中进行的单倍群一致率分析表明, GSA、CGA 与 1240K 面板在 306 对一级亲缘关系中均表现出极高的一致性, 能够准确地将所有亲属样本划分至同一单倍群。特别地, GSA 面板在 109 例中国汉族样本构成的 384 对 1~12 级亲缘关系中展现出卓越的分类稳定性: 即使亲缘关系跨越多个世代, 所有样本仍被一致地划分至相同单倍群, 说明该面板在复杂亲

缘结构下仍具有较高的系统发育分辨能力。然而, 需指出的是, 本研究中 GSA 面板的验证样本完全来源于中国汉族家系。尽管该群体具有相对较高的遗传多样性和代表性, 但仍不能完全代表全球人群的 Y 染色体谱系结构多样性。因此, 当前基于汉族样本所表现出的优异性能是否适用于其他地理人群(如非洲、欧洲或土著群体), 仍需进一步验证。尤其在某些具有独特 Y 谱系结构的群体中, 若 GSA 面板未覆盖其特有分支或私有突变, 可能导致分辨率下降或误判风险升高。未来研究应在更广泛的人群中系统评估该面板的跨人群适用性, 包括纳入更多具有独特 Y 染色体谱系的族群(如撒哈拉以南非洲人群、大洋洲人群或高加索人群等), 从而全面检验其全球普适性。此外, 建议结合 Y-STR 分型与 Y-SNP 分型进行混合分析^[49], 以进一步提升中远亲缘关系的分辨能力, 并推动 Y 染色体标记在法医遗传学和人类谱系研究中的广泛应用。

3.3 不同量级位点组合下单倍群一致性分析

基于 WGS 的不同量级 Y-SNP 组合研究显示, Y-SNP 位点数量与单倍群分类一致性之间存在显著的正相关关系, 特别是在位点数由 1 000 增加至 10 000 的区间内, 一致性显著提升。随着位点数超过 10 000, 该提升趋势趋于平稳, 表明在 WGS 数据中, 当与 ISOGG 谱系树重叠 Y-SNP 覆盖达到一定饱和度后, 新增位点对单倍群分类一致性的贡献逐渐减弱。进一步分析表明, 基于 GSA 与 CGA 位点集随机筛选出的 Y-SNPs 组合, 其单倍群分类错判率随位点数量增加而逐渐降低。当位点数量处于 1 000~5 000 区间时, 不同 Y-SNP 组合对单倍群判定的准确性存在显著差异。当 Y-SNP 数量大于 5 000 后单倍群判别一致性较高并趋于稳定, 然而本研究在位点选择中采取的是随机抽样的方法, 可能导致结果无法反应实际应用中核心位点的作用, 在实际筛选位点时应优先保留 ISOGG 谱系树的关键分支位点。在资源受限或大规模人群研究中, 合理优化 Y-SNP 位点选择策略, 优先覆盖具有谱系分辨能力的核心位点, 有助于在降低测序或分型成本的同时, 保持较高的单倍群分类准确性。这一策略对于考古学、基因组学研究及法医学应用具有重要的实践意义。

4 结 论

在进行单倍群精细程度划分以及人群区分方面, GSA 与 CGA 位点集在单倍群判别上兼具高一

致性和高分辨率,更接近于WGS所分类的单倍群。分类一致性在WGS数据中随机挑选的Y-SNPs(1 000~10 000个标记)范围内呈正相关,但当标记数量超过10 000时,其分类一致性趋于平缓。所有家系样本中4组位点集均未出现单倍群不匹配亲缘关系对。本研究从不同角度对不同位点集数据进行评估,可为应用高密度Y-SNP位点集和不同量级Y-SNP位点组合的祖先溯源和法医学应用提供参考。

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

PIBB_20250481_Figure_S1.tif

PIBB_20250481_Table1_S1.xlsx

PIBB_20250481_Table1_S2.xlsx

PIBB_20250481_Table1_S3.xlsx

参 考 文 献

- [1] Hallast P, Ebert P, Loftus M, *et al.* Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature*, 2023, **621**(7978): 355-364
- [2] Navarro-López B, Baeta M, Moreno-López O, *et al.* Y-chromosome analysis recapitulates key events of Mediterranean populations. *Heliyon*, 2024, **10**(16): e35329
- [3] Qian X, Hou J, Wang Z, *et al.* Next generation sequencing plus (NGS+) with Y-chromosomal markers for forensic pedigree searches. *Sci Rep*, 2017, **7**(1): 11324
- [4] Jangravi Z, Alikhani M, Arefnezhad B, *et al.* A fresh look at the male-specific region of the human Y chromosome. *J Proteome Res*, 2013, **12**(1): 6-22
- [5] Mahal D G, Matsoukas I G. The geographic origins of ethnic groups in the Indian subcontinent: exploring ancient footprints with Y-DNA haplogroups. *Front Genet*, 2018, **9**: 4
- [6] Pinotti T, Bergström A, Geppert M, *et al.* Y Chromosome Sequences Reveal a short Beringian standstill, rapid expansion, and early population structure of Native American Founders. *Curr Biol*, 2019, **29**(1): 149-157.e3
- [7] Kayser M, Branicki W, Parson W, *et al.* Recent advances in forensic DNA phenotyping of appearance, ancestry and age. *Forensic Sci Int Genet*, 2023, **65**: 102870
- [8] Colaco S, Modi D. Genetics of the human Y chromosome and its association with male infertility. *Reprod Biol Endocrinol*, 2018, **16**(1): 14
- [9] Zhou Z, Li Z, Yao Y, *et al.* Validation of phylogenetic informative Y-InDels in Y-chromosomal haplogroup O-M175. *Front Genet*, 2023, **14**: 1182028
- [10] Calafell F, Larmuseau M H D. The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. *Hum Genet*, 2017, **136**(5): 559-573
- [11] Naidoo T, Xu J, Vicente M, *et al.* Y-chromosome variation in southern African khoe-San populations based on whole-genome sequences. *Genome Biol Evol*, 2020, **12**(7): 1031-1039
- [12] Barbieri C, Hübner A, Macholdt E, *et al.* Refining the Y chromosome phylogeny with southern African sequences. *Hum Genet*, 2016, **135**(5): 541-553
- [13] Tao R, Li M, Chai S, *et al.* Developmental validation of a 381 Y-chromosome SNP panel for haplogroup analysis in the Chinese populations. *Forensic Sci Int Genet*, 2023, **62**: 102803
- [14] Navarro-López B, Granizo-Rodríguez E, Palencia-Madrid L, *et al.* Phylogeographic review of Y chromosome haplogroups in Europe. *Int J Legal Med*, 2021, **135**(5): 1675-1684
- [15] Rowold D J, Chennakrishnaiah S, Gayden T, *et al.* The Y-chromosome of the Soliga, an ancient forest-dwelling tribe of South India. *Gene*, 2020, **763S**: 100026
- [16] Zerjal T, Xue Y, Bertorelle G, *et al.* The genetic legacy of the mongols. *Am J Hum Genet*, 2003, **72**(3): 717-721
- [17] Lalueza-Fox C, Gigli E, Bini C, *et al.* Genetic analysis of the presumptive blood from Louis XVI, King of France. *Forensic Sci Int Genet*, 2011, **5**(5): 459-463
- [18] Charlier P, Olalde I, Solé N, *et al.* Genetic comparison of the head of Henri IV and the presumptive blood from Louis XVI (both Kings of France). *Forensic Sci Int*, 2013, **226**(1/2/3): 38-40
- [19] Larmuseau M H D, Delorme P, Germain P, *et al.* Genetic genealogy reveals true Y haplogroup of House of Bourbon contradicting recent identification of the presumed remains of two French Kings. *Eur J Hum Genet*, 2014, **22**(5): 681-687
- [20] Wang C, Yan S, Hou Z, *et al.* Present Y chromosomes reveal the ancestry of Emperor CAO Cao of 1800 years ago. *J Hum Genet*, 2012, **57**(3): 216-218
- [21] Wang C C, Yan S, Yao C, *et al.* Ancient DNA of Emperor CAO Cao's granduncle matches those of his present descendants: a commentary on present Y chromosomes reveal the ancestry of Emperor CAO Cao of 1800 years ago. *J Hum Genet*, 2013, **58**(4): 238-239
- [22] Ardui S, Ameer A, Vermeesch J R, *et al.* Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res*, 2018, **46**(5): 2159-2168
- [23] Logsdon G A, Vollger M R, Eichler E E. Long-read human genome sequencing and its applications. *Nat Rev Genet*, 2020, **21**(10): 597-614
- [24] Byrška-Bishop M, Evani U S, Zhao X, *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, 2022, **185**(18): 3426-3440.e19
- [25] Anderson K, Cañadas-Garre M, Chambers R, *et al.* The challenges of chromosome Y analysis and the implications for chronic kidney disease. *Front Genet*, 2019, **10**: 781
- [26] Alvarez-Cubero M J, Santiago O, Martínez-Labarga C, *et al.* Methodology for Y chromosome capture: a complete genome sequence of Y chromosome using flow cytometry, laser microdissection and magnetic streptavidin-beads. *Sci Rep*, 2018, **8**(1): 9436
- [27] Rohrlach A B, Papac L, Childebayeva A, *et al.* Using Y-chromosome capture enrichment to resolve haplogroup H₂ shows new evidence for a two-path Neolithic expansion to Western

- Europe. *Sci Rep*, 2021, **11**(1): 15005
- [28] García-Fernández C, Lizano E, Telford M, *et al.* Y-chromosome target enrichment reveals rapid expansion of haplogroup R1b-DF27 in Iberia during the Bronze Age transition. *Sci Rep*, 2022, **12**(1): 20708
- [29] 23mofang. 23MOFANG WHITE PAPER 1-1. Chengdu: 23mofang, 2017[2025-06-30]. <https://www.23mofang.com/white-papers/chip/>
- [30] Mallick S, Micco A, Mah M, *et al.* The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. *Sci Data*, 2024, **11**(1): 182
- [31] Ralf A, van Oven M, Montiel González D, *et al.* Forensic Y-SNP analysis beyond SNaPshot: high-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing. *Forensic Sci Int Genet*, 2019, **41**: 93-106
- [32] Claerhout S, Verstraete P, Warnez L, *et al.* CSYseq: the first Y-chromosome sequencing tool typing a large number of Y-SNPs and Y-STRs to unravel worldwide human population genetics. *PLoS Genet*, 2021, **17**(9): e1009758
- [33] Liu J, Jiang L, Zhao M, *et al.* Development and validation of a custom panel including 256 Y-SNPs for Chinese Y-chromosomal haplogroups dissection. *Forensic Sci Int Genet*, 2022, **61**: 102786
- [34] He G, Wang M, Miao L, *et al.* Multiple founding paternal lineages inferred from the newly-developed 639-plex Y-SNP panel suggested the complex admixture and migration history of Chinese people. *Hum Genomics*, 2023, **17**(1): 29
- [35] 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
- [36] Danecek P, Auton A, Abecasis G, *et al.* The variant call format and VCFtools. *Bioinformatics*, 2011, **27**(15): 2156-2158
- [37] Genovese G, Rockweiler N B, Gorman B R, *et al.* BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. *Bioinformatics*, 2024, **40**(2): btae038
- [38] Chen H, Lu Y, Lu D, *et al.* Y-LineageTracker: a high-throughput analysis framework for Y-chromosomal next-generation sequencing data. *BMC Bioinformatics*, 2021, **22**(1): 114
- [39] Köksal Z, Börsting C, Bailliet G, *et al.* Application of targeted Y-chromosomal capture enrichment to increase the resolution of native American haplogroup Q. *Hum Mutat*, 2024, **2024**: 3046495
- [40] Wang M, Sun Q, Feng Y, *et al.* Paleolithic divergence and multiple Neolithic expansions of ancestral nomadic emperor-related paternal lineages. *J Genet Genomics*, 2025, **52**(4): 502-512
- [41] Francalacci P, Morelli L, Angius A, *et al.* Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science*, 2013, **341**(6145): 565-569
- [42] Dzehverovic M, Jusic B, Pilav A, *et al.* Kinship analysis of skeletal remains from the Middle Ages. *Forensic Sci Int Genet*, 2023, **63**: 102829
- [43] Jaisamut K, Rasmeepaisarn K, Pitiwararom R, *et al.* Paternal genetic landscape of contemporary Thai populations in the borderland provinces of Thailand and Myanmar. *Sci Rep*, 2025, **15**(1): 6300
- [44] Babić Jordamović N, Kojović T, Dogan S, *et al.* Haplogroup prediction using Y-chromosomal short tandem repeats in the general population of Bosnia and Herzegovina. *Front Genet*, 2021, **12**: 671467
- [45] Song M, Zhou Y, Zhao C, *et al.* YHP: Y-chromosome Haplogroup Predictor for predicting male lineages based on Y-STRs. *Forensic Sci Int*, 2024, **361**: 112113
- [46] Narang A, Uppilli B, Vivekanand A, *et al.* Frequency spectrum of rare and clinically relevant markers in multiethnic Indian populations (ClinIndb): a resource for genomic medicine in India. *Hum Mutat*, 2020, **41**(11): 1833-1847
- [47] Cherukuri PF, Soe MM, Condon DE, *et al.* Establishing analytical validity of BeadChip array genotype data by comparison to whole-genome sequence and standard benchmark datasets. *BMC Med Genomics*, 2022, **15**(1): 56
- [48] Cao Y, Li L, Xu M, *et al.* The ChinaMAP analytics of deep whole genome sequences in 10, 588 individuals. *Cell Res*, 2020, **30**(9): 717-731
- [49] Song M, Wang Z, Lyu Q, *et al.* Paternal genetic structure of the Qiang ethnic group in China revealed by high-resolution Y-chromosome STRs and SNPs. *Forensic Sci Int Genet*, 2022, **61**: 102774
- [50] Roots S, Behar D M, Järve M, *et al.* Phylogenetic applications of whole Y-chromosome sequences and the Near Eastern origin of Ashkenazi Levites. *Nat Commun*, 2013, **4**: 2928
- [51] Platt D E, Artinian H, Mouzaya F, *et al.* Autosomal genetics and Y-chromosome haplogroup L1b-M317 reveal Mount Lebanon Maronites as a persistently non-emigrating population. *Eur J Hum Genet*, 2021, **29**(4): 581-592
- [52] Li W, Xiong Z, He K, *et al.* TGDD: a genomics-based database for tumor treatment drugs. *iCell*, 2024, **1**(1): UYBT8671

Assessing High-density Y-SNP Panels for Paternal Haplogroup Assignment in Forensic Practice*

ZHANG De-Qin^{1,2)}, WANG Chun-Nian^{2,3)}, LOU Lin-Lin^{2,4)}, NI Meng⁵⁾, GAO Jing^{2,3)},
HUANG Jiang^{1)**}, JIANG Li^{2)**}

⁽¹⁾Toxicity Testing Center, Department of Forensic Medicine, Guizhou Medical University, Guiyang 561113, China;

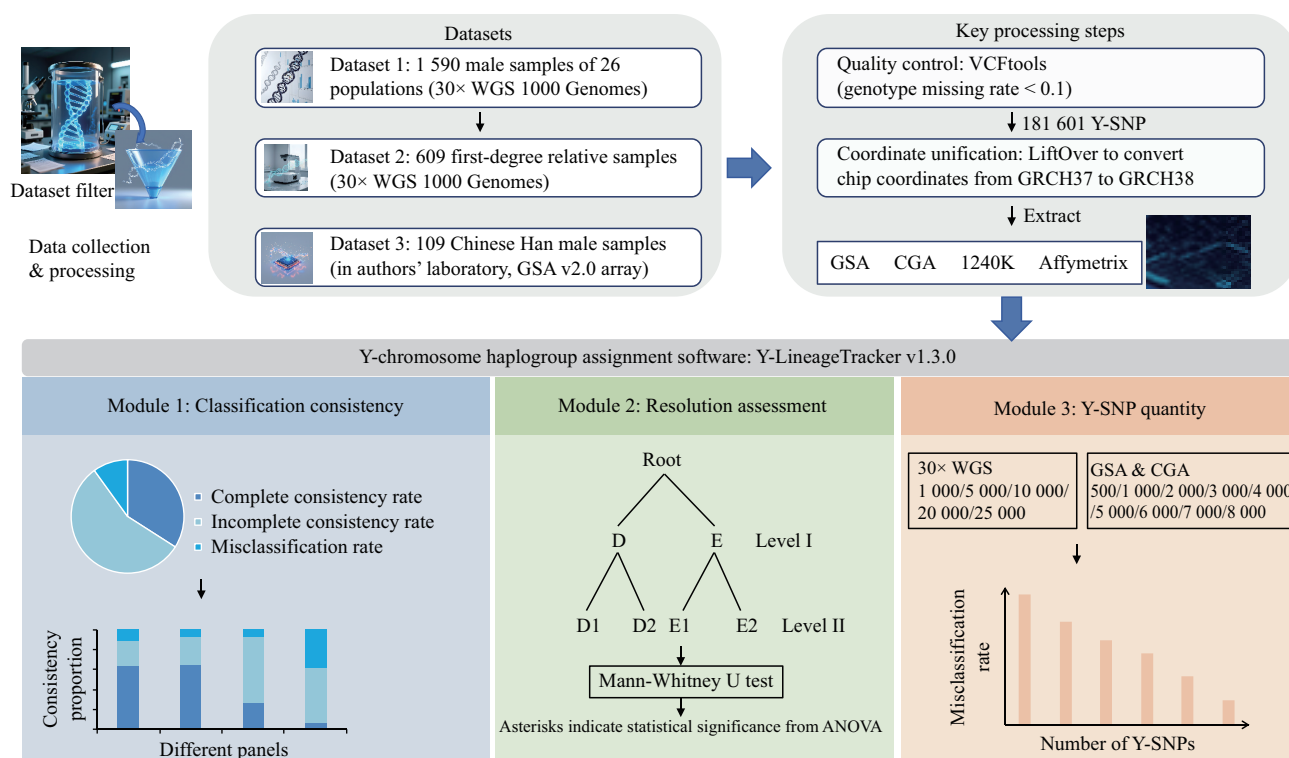
²⁾Institute of Forensic Science, Key Laboratory of Forensic Genetics, Beijing Engineering Research Center of Crime Scene Evidence Examination, National Engineering Laboratory for Forensic Science, Beijing 100038, China;

³⁾School of Investigation, People's Public Security University of China, Beijing 100038, China;

⁴⁾School of Forensic Medicine, Shanxi Medical University, Jinzhong 030600, China;

⁵⁾Jiangsu Key Laboratory of Phylogenomics and Comparative Genomics, School of Life Sciences, Jiangsu Normal University, Xuzhou 221116, China)

Graphical abstract



* This work was supported by grants from the National Key Research and Development Program of China (2022YFC3341004), The National Natural Science Foundation of China (81772027), and the Science and Technology Program of the Ministry of Public Security (2023JSZ02).

** Corresponding author.

HUANG Jiang. Tel: 86-851-88596769, E-mail: mmm_hj@126.com

JIANG Li. Tel: 86-10-83752707, E-mail: jl@mail.bnu.edu.cn

Received: November 3, 2025 Accepted: December 26, 2025

Abstract Objective The accuracy of Y-chromosome haplogroup assignment is crucial for tracing paternal lineage in male samples. With the advancement of high-throughput sequencing technologies, high-density Y-SNP genotyping from whole-genome or array-based data has become a standard method for determining Y-chromosome haplogroups. This study systematically evaluated the performance of 4 commonly used high-density SNP genotyping systems—namely, the Global Screening Array (GSA), Chinese Genotyping Array (CGA), Affymetrix array, and the 1240K capture panel—for haplogroup assignment. This work provides a reference for data comparison across different systems. **Methods** We extracted genotype data for the 4 Y-SNP panels from 30× whole-genome sequencing (WGS) data of 1 590 male samples from the 1000 Genomes Project. Additionally, GSA array genotype data from 384 relative pairs (spanning 1st- to 12th-degree relationships) from 109 Chinese Han families were collected. Haplogroup assignment was performed using Y-LineageTracker v1.3.0 software. We assessed the concordance and resolution of haplogroup assignments between the four Y-SNP panels and the WGS data. The consistency and resolution of haplogroup assignments were also evaluated for both the 1000 Genomes Project samples and the 109 family samples collected in this study. Furthermore, the impact of varying numbers of Y-SNPs on haplogroup assignment was examined. **Results** The GSA and CGA panels demonstrated superior resolution and discrimination of haplogroup subclades compared with the other two panels. The haplogroup assignments from the GSA, CGA, and 1240K panels showed high concordance with WGS data, with consistency rates exceeding 88.70%, whereas the Affymetrix platform exhibited a significantly lower consistency rate of 61.89%. Specifically, the GSA and CGA panels consistently demonstrated superior performance compared with the other two panels in the assignment of haplogroups O-M175 and H-L901, achieving complete concordance (100%) for both haplogroups. In contrast, the Affymetrix panel erroneously assigned all individuals belonging to haplogroup O-M175 to haplogroup K2-M526. Furthermore, its accuracy for haplogroup H-L901 was exceedingly low, at merely 1.41%. This poor performance was characterized by the misassignment of 98.59% of H-L901 samples—specifically, 1.41% to J-M304 and a predominant 97.18% to F-M89. For haplogroup R-M207, all four panels exhibited uniformly high levels of consistency, with concordance values exceeding 94.00%. Notably, for haplogroup E-M96, the 1240K and Affymetrix panels outperformed the GSA and CGA panels in terms of concordance, representing the first instance in which these two panels surpassed the latter. Conversely, for haplogroups J-M304, Q-M242, and I-M170, all 4 panels showed relatively elevated misclassification rates, with the Affymetrix array demonstrating the poorest overall performance. None of the four panels showed any discordant haplogroup assignments among the familial relative pairs analyzed. A positive correlation was observed between the number of Y-SNPs (ranging from 1 000 to 10 000) and classification consistency; however, classification consistency plateaued when the number of Y-SNPs exceeded 10 000. Furthermore, a random sampling analysis conducted on the GSA and CGA panels demonstrated that the haplogroup misclassification rate exhibited negligible fluctuation across the Y-SNP range of 500 to 1 000. Conversely, a marked enhancement in classification consistency was observed as the number of markers increased from 1 000 to 5 000, ultimately reaching a plateau within the interval of 5 000 to 8 000 markers. **Conclusion** These findings indicate that the GSA and CGA panels provide high resolution and concordance, delivering reliable Y-haplogroup assignment for forensic investigations.

Key words Y-SNP, NRY haplogroup classification, paternal genealogy, ancestry inference

DOI: 10.3724/j.pibb.2025.0481

CSTR: 32369.14.pibb.20250481