 DOI: https://doi.org/10.12345/yzlcyxzz.v8i9.31988

Forensic Characteristics and Application Efficiency of Shenzhen Han Chinese via 131 STRs Based on MPS Platform

Gaoyuan Song Dongyang Qu Cheng Xu Xiangqin Li*

Criminal Investigation Department of Shenzhen Municipal Public Security Burreau, Shenzhen, Guangdong, 518000, China

Abstract

Objective Based on MPS technology, 51 autosomal short tandem repeats and 80 Y chromosome short tandem repeats genetic markers were sequenced in the Shenzhen Han Chinese, aiming to investigate the capabilities of it in enhancing the effectiveness of forensic medicine so as to provide more comprehensive and precise data support for individual identification and complex kinship identification. **Methods** The Forensic Analysis System Multiplecues SetB Kit on the MGISEQ-2000 platform sequenced 131 STRs of 468 unrelated individuals of Shenzhen Han, and the forensic parameters of A-STRs and Y-STRs were calculated based on length and sequence polymorphism, respectively. **Results** Based on length and sequence polymorphism, for the 51 A-STRs, 535 and 980 different alleles were detected, respectively; for the 80 Y-STRs, 530 and 816 different alleles were detected, respectively. Meanwhile, the CPE were 1-3.146474E-20 and 1-2.953016E-23; the CPD were 1-6.428174E-58 and 1-2.411721E-64 based on the length and sequence polymorphism, respectively. Among the 80 Y-STRs, based on length and sequence polymorphism levels, HD was 0.99990187 and 0.99994648; HMP was 0.00220762 and 0.00216311. **Conclusion** The sequence polymorphism of the 131 STRs based on the MPS platform significantly improved the genetic polymorphism of the STRs .

Keywords

MPS; Shenzhen Han population; Sequence polymorphisms

基于 MPS 的 131 个 STRs 在深圳汉族中的法医学应用

宋高原 曲冬阳 徐程 李湘秦*

深圳市公安局刑事警察支队，中国 · 广东 深圳 518000

摘 要

目的 基于MPS技术对深圳汉族群体的51个A-STRs和80个Y-STRs遗传标记进行测序，探索其在法医个体识别和复杂亲属关系鉴定中的作用。**方法** 使用Forensic Analysis System Multiplecues SetB Kit在MGISEQ—2000平台上对468例深圳汉族男性无关个体的131个STRs遗传标记进行测序，并分别统计长度和序列多态性法医学参数。**结果** 基于长度和序列多态性，51个A—STRs基因座，分别检出535种和980种不同的等位基因，80个Y—STRs基因座，分别检出530种和816种不同的等位基因。在长度和序列多态性水平上，CPE值分别为1—3.146474E—20和1—2.953016E-23，CPD值分别为1—6.428174E—58和1—2.411721E-64。80个Y-STR基因座中，基于长度和序列多态性水平，HD分别为0.99990187和0.99994648，HMP分别为0.00220762和0.00216311。**结论** 基于MPS平台获得的131个STR遗序列多态性信息，提高了STR遗传标记的遗传多态性和法医学应用效能。

关键词

MPS；深圳汉族；序列多态性

1 引言

大规模平行测序（Massively parallel sequencing, MPS），也称新一代测序技术（Next generation sequencing, NGS），具有高通量、低成本和高效率的特点，能够在一次测序中并行测序分析上百万个 DNA 片段，同时检测多种 STR 遗传标记类型 [1, 2]。与传统的 CE 测序平台相比，MPS 可以同时获得 STR 的长度及序列多态性信息 [2]。

本研究基于 MPS 平台，采用 Forensic Analysis System Multiplecues SetB Kit（51 个 A-STR 和 80 个 Y-STR）对 468 名深圳男性汉族无关个体进行测序，基于长度和序列多态性的 STRs 分型数据，探究其在法医学鉴定效能和群体适用性方面的差异。本研究旨在探索 MPS 平台和多遗传标记 SetB 试剂盒在提升法医学应用效能方面的潜力，为法医个体识别和复杂亲属关系鉴定等提供更为丰富和准确的数据支持。

【作者简介】宋高原（1988–），男，中国河南周口人，博士，从事法医物证检验鉴定研究。

【通讯作者】李湘秦（1975–），男，中国新疆伊犁人，硕士，从事法医物证检验鉴定研究。

2 材料与方法

2.1 样本 DNA 提取、文库构建及测序

遵循知情同意原则，使用 FTA 卡（博坤生物科技有限公司，长春）随机采集广东深圳地区 468 例男性汉族无关个体血液样本。随后，根据 FAN 等方法进行了 DNA 提取及文库的构建 [3]。在 MGISEQ-2000 平台（深圳华大智造科技股份有限公司）上进行单端 400 nt 测序，并通过调整配置的 STRait Razor 3.0 实现 STR 的调用。

2.2 统计学分析

在研究中，我们首先使用 Genepop 4.7 软件 [4]，对 A-STR 的长度及序列多态性水平的群体分型数据进行了哈迪 - 温伯格平衡（Hardy–Weinberg equilibrium, HWE）检验，为了确保结果的准确性，我们还对 p 值进行了邦费罗尼（Bonferroni）校正。然后，运用 STRsAF 2.1.5 软件 [5]，分别基于序列多态性和长度多态性，对 131 个 STR 基因座进行了法医学相关参数的计算和统计，包括期望杂合度（expected heterozygosity, H_{exp}）、观测杂合度（observed heterozygosity, H_{obs}）、鉴别能力（power of discrimination, PD）、排除概率（exclusion probability, PE）、匹配概率（match probability, PM）、典型父系指数（typical paternity index, TPI）、多态性信息含量（polymorphism information content, PIC）等。此外，我们还采用直接计数法统计了 80 个 Y-STR 基因座在深圳汉族群体中的单倍型及单倍型频率；并根据公式，计算了 Y-STRs 的单倍型差异度（Haplotype Diversity，HD）、单倍型匹配概率（Haplotype Match Probability，HMP）和分辨能力（haplotype discrimination capacity，DC）等参数。以上参数计算公式为：$PM=\sum \chi^2$，$PD=1-\sum \chi^2$，$HD=N(1-\sum \chi^2)/N-1$（其中 χ 是每个 Y-STR 单倍型的频率，N 为样本数）。

3 结果

3.1 STRs 长度多态性的法医学相关参数

经 Bonferroni 校正后（$p > 0.05/51$），51 个 A-STR 与 80 个 Y-STR 基因座的长度多态性及序列多态性均符合 HWE 检验。51 个 A-STR 共检出 535 种不同的等位基因，最少 STR 的有 6 种，分别是 D4S2408、D6S474、D3S4529、D1GATA113 及 TH01，而最多是 FGA 有 21 种。通过统计分析发现 51 个 A-STR 的 GD 值在 0.605（TPOX）—0.879（D7S3048）之间，H_{obs} 值在 0.612（TPOX）—0.886（D6S1043），平均值分别为 0.780（±0.00977），H_{exp} 值为 0.604（TPOX）—0.878（D7S3048），平均值为 0.783（±0.0092）。PIC 值 0.544（TPOX）—0.864（D7S3048），PE 值为 0.305（TPOX）—0.767（D6S1043），联合 PE 值为 1-3.146474E-20，PD 值 0.782（TPOX）—0.970（D7S3048），联合 PD 值分别为 1—6.428174E-58。PM 值范围从 0.030（D7S3048）—0.218（TPOX）（如图 1 所示）。最大的 TPI 值出现在 D6S1043 基因座上，为 4.389。

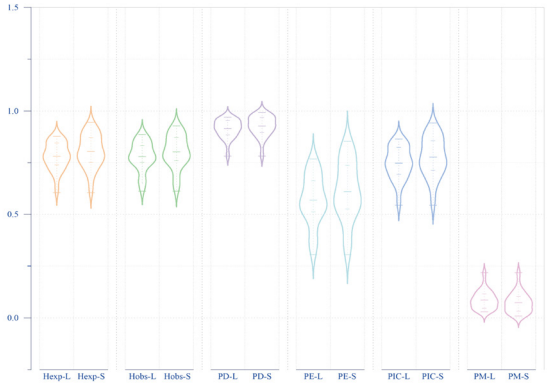

图 1 51 个 A–STRs 基因座的法医学参数
（L 代表长度多态性，S 代表序列多态性）

80 个 Y-STR 基因座在长度多态性水平上检出 530 种不同的等位基因，最多的是 DYS518 有 21 种。通过统计分析发现，除了基因座 DYS502 的 GD 值为 0 外，其他 Y-STR 基因座的 GD 值在 0.008（DYS472）和 0.951（双拷贝基因座 DYF387S1a/b）之间波动，HD、HMP 和 DC 分别为 0.99990187、0.00220762 和 0.97679325。

3.2 STRs 序列多态性的法医学相关参数

在 51 个 A-STR 基因座中，基于序列多态性共检出 980 种不同的等位基因，最少的 STR 只检出有 6 种，分别是 D1GATA113 和 TH01，而最多的是 D7S1517 共检出 56 个等位基因位点。GD 值在 TPOX 基因座的 0.605 与 D7S1517 基因座的 0.946 之间变化，H_{obs} 值范围在 0.612（TPOX）至 0.928（D7S3048）之间，平均值为 0.80169（±0.01125）。H_{exp} 值为 0.604（TPOX）—0.945（D7S1517），平均值为 0.804（±0.01122）。PIC 值为 0.544（TPOX）—0.942（D7S1517），PE 值为 0.305（TPOX）—0.853（D7S3048），联合 PE 值为 1-2.953016E-23，PD 值 0.782（TPOX）—0.992（D7S3048），联合 PD 值为 1—2.411721E—64，PM 值在 0.008（D7S1517）到 0.218（TPOX）之间（如图 1 所示）。最大的 TPI 值出现在 D7S3048 基因座上，为 6.971。

80 个 Y-STR 基因座的序列多态性水平进行统计分析，共检出 816 种不同的等位基因，最多的是 DYS449，有 62 种。最小和最大 GD 值分别出现在 DYS472 和 DYF387S1a/b 基因座上，分别为 0.008 和 0.983。其 HD、HMP、DC 分别为 0.99994648、0.00216311、0.98734177。

3.3 STRs 高频等位基因的序列多态性

40 个 A-STR 的 189 个等位基因共检出 444 个亚分型，其频率均大于 0.001，D8S1132 在等位基因 20 和 22 上亚分型最多，分别有 9 种序列多态性分型。13 个公安部要求的常染色体核心基因座也发现序列多态性亚分型，分别是 D3S1358、D13S317、D7S820、D16S539、D2S1338、CSFIPO、VWA、D21S11、D6S1043、D8S1179、D5S818、D12S391 和 FGA。24 个 A-STR 的 83 个等位基因共检出

117 个亚分型，其频率大于 0.01, 包括 5 个核心基因座，分别是 D3S1358、D2S1338、D21S11、D8S1179、D12S391（部分位点如表 1 所示），亚分型出现频率最高的是 D21S11 的 29B，亚分型频率为 0.089（如表 1 所示）。

a 基于片段长度多态性获得的等位基因；b 基于重复区域序列多态性获得的等位基因。

30 个 Y-STR 的 126 个等位基因共检出 284 个亚分型，其频率均大于 0.001, 其中包括 12 个公安部要求的核心 Y-STR 基因座，分别是 DYS385a/b、DYS389I/II、DYS390、DYS392、DYS393、DYS437、DYS438、DYS448、DYS456、DYS458、DYS635、DYS533，14 个 Y-STR 的 59 个等位基因共检出 97 个亚分型，其频率大于 0.01，包括 6 个核心 Y-STR 位点，分别是 DYS389II、DYS390、DYS437、DYS438、DYS448、DYS635（部分位点如表 1），亚分型出现频率最高的是 DYS437 的 14B，亚分型频率为 0.186（如表 1 所示）。

表 1 部分 STR 基因座基于长度多态性和重复区域序列的等位基因比较（n=468，F ＞ 0.01）

STR	等位基因 [a]	等位基因 [b]	STR 重复序列多态性	F
D21S11	29	29A	[TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10	0.143
		29B	[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	0.089
		29C	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10	0.015
	30	30A	[TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	0.110
		30B	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	0.074
		30C	[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]12	0.035
		30D	[TCTA]7[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10	0.027
D3S1358	16	16A	[TCTA]1[TCTG]2[TCTA]13	0.250
		16B	[TCTA]1[TCTG]3[TCTA]12	0.074
	17	17A	[TCTA]1[TCTG]2[TCTA]14	0.139
		17B	[TCTA]1[TCTG]3[TCTA]13	0.079
D8S1179	13	13A	[TCTA]1[TCTG]1[TCTA]11	0.123
		13B	[TCTA]13	0.077
	14	14A	[TCTA]1[TCTG]1[TCTA]12	0.096
		14B	[TCTA]2[TCTG]1[TCTA]11	0.042
		14C	[TCTA]14	0.032
DYS389II	29	29A	[TAGA]9[CAGA]3N48[TAGA]13[CAGA]4	0.179
		29B	[TAGA]10[CAGA]3N48[TAGA]12[CAGA]4	0.108
		29C	[TAGA]11[CAGA]3N48[TAGA]11[CAGA]4	0.011
	30	30A	[TAGA]10[CAGA]3N48[TAGA]13[CAGA]4	0.068
		30B	[TAGA]11[CAGA]3N48[TAGA]12[CAGA]4	0.059
		30C	[TAGA]9[CAGA]3N48[TAGA]14[CAGA]4	0.038
		30D	[TAGA]11[CAGA]3N48[TAGA]11[CAGA]5	0.017
		30E	[TAGA]10[CAGA]3N48[TAGA]12[CAGA]5	0.011
DYS437	14	14A	[TCTA]8[TCTG]2[TCTA]4	0.435
		14B	[TCTA]9[TCTG]1[TCTA]4	0.186
DYS447	24	24A	[TTATA]8[TTATT]1[TTATA]7[TTATT]1[TTATA]7	0.118
		24B	[TTATA]7[TTATT]1[TTATA]7[TTATT]1[TTATA]8	0.061
		24C	[TTATA]7[TTATT]1[TTATA]8[TTATT]1[TTATA]7	0.036
DYS448	20	20A	[AGAGAT]11N42[AGAGAT]9	0.169
		20B	[AGAGAT]12N42[AGAGAT]8	0.063

4 讨论

在本研究中，我们对深圳地区 468 名汉族无关个体的 131 个 STR 遗传标记进行了全面分析，以探讨其法医学参数及特征。经 Bonferroni 校正后，所有 51 个 A-STR 基因座均符合 Hardy-Weinberg（HWE）平衡，所有 A-STR 基因座的 H_{obs} 和 H_{exp} 均大于 0.5，具有较高的杂合度，对法医学个人识别具有重要价值。

4.1 STR 序列多态性显著高于长度多态性

51 个 A-STR 基因座中，基于长度多态性共检出 535 种不同的等位基因，基于序列多态性共检出 980 种不同的等位基因，多态性增加约 83%。80 个 Y-STR 基因座在长度多态性水平上检出 530 个等位基因，在序列多态性上共检出 816 个不同的等位基因，多态性增加约 54%。40 个 A-STR 和

30 个 Y-STR 基于序列多态性的 GD 值显著高于基于长度多态性的 GD 值。另外，我们发现 14 个 A-STR（D12S391、D7S1517、D7S3048、D8S1132、D21S11、D13S325、D2S1338、D22GATA198B05、D11S2368、D3S1358、D11S4463、D9S1122、D4S2366 和 vWA） 和 10 个 Y-STR（DYS447、DYS449、DYS389II、DYS518、DYS527a/b、DYS390、DYS635、DYS448、DYS437 和 DYS552）的序列等位基因数比长度等位基因数增加 1-4 倍（如表 1 所示），这与之前部分研究结果一致 [6-8]。这些充分说明 STR 的序列多样性更丰富，可为复杂亲缘关系的鉴定提供支撑。

4.2 STR 序列多态性集中在基因座高频等位基因上

40 个 A-STR 的序列多态性均集中在高频等位基因，D21S391 的等位基因序列多态性共检出 42 个亚分型，其集中分布在长度为 16-26 的高频等位基因位点上，类似的 STR 还有 D2S1338、D3S1358、vWA 等基因座分别检出 30、10、8 个序列多态性亚分型，均全部分布在高频等位基因位点上。同时，我们在 Y-STR 的序列多态性也发现类似现象，如 DYS449 和 DYS389II 等的等位基因分别检出 48 和 22 个序列多态性亚分型，均集中分布在高频等位基因位点上（如表 1 所示）。

4.3 A–STR 亚分型可有效缩小单亲比中范围

在单亲认证以及日常案件单亲比对中，经常出现单亲比中信息数多达百条甚至超千条，增加信息研判与侦查的难度，通过 A-STR 序列多态性信息联合应用可以有效缩小比中范围，如表 2 中所示，D21S11 的等位基因 29 的亚分型 29A、29B 和 29C 出现的频率分别为 0.143、0.089 和 0.015；D3S1358 的等位基因 17 的亚分型 17A、18B 出现的频率分别为 0.139 和 0.079；D8S1179 的等位基因 14 的亚分型 14A、14B、14C 出现的频率分别为 0.096、0.042、0.032。另外，通过本次测序发现，PentaD、TPOX、TH01、D19S433、D18S51、D10S1248、D1GATA113、D20S470、D22S1045、D3S3045、D6S1017 的长度多态性与序列多态性一致，在与一代 STR 数据比对中，这些位点可以用于筛选优先对象。

4.4 Y–STR 亚分型为人员家系排查提供精确信息

伴随着 Y 库建设持续进行，不少单位通过 Y-STR 找家系进而破获一些命案要案 [9-12]，但在 Y 家系排查中也经常出现比中人员较多，来自不同的家系，且分布范围广等问题，给侦查排查带来较大困难。通过 Y-STR 序列多态性可以对长度比中信息进行有效甄选，例如 DYS389II 的等位基因 30，有 5 种亚分型 30A、30B、30C、30D 和 30E，其频率分别为 0.068、0.059、0.038、0.017、0.011，DYS447 的等位基因 24，有 3 种亚分型，分别是 24A、24B 和 24C，其频率分别为 0.118、0.061 和 0.036（如表 2 所示）等，极大的缩小家系范围，精确锁定嫌疑人所在家系，提高办案侦查效率。

综上所述，通过 MPS 获取的 131 个 STR 序列信息对于群体遗传学研究具有重要的应用价值，为深圳汉族群体提供一种具有科学可行性的个体识别和复杂亲缘关系鉴定方法。另外，由于实验样本有限，其他位点是否存在序列多态性，还需要进一步验证。

参考文献

[1] ZHONG Y, XU F, WU J, et al. Application of Next Generation Sequencing in Laboratory Medicine [J]. Ann Lab Med, 2021, 41(1): 25-43.

[2] HU T, CHITNIS N, MONOS D, et al. Next-generation sequencing technologies: An overview [J]. Hum Immunol, 2021, 82(11): 801-11.

[3] FAN H, WANG L, LIU C, et al. Development and validation of a novel 133-plex forensic STR panel (52 STRs and 81 Y-STRs) using single-end 400 bp massive parallel sequencing [J]. Int J Legal Med, 2022, 136(2): 447-64.

[4] ROUSSET F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux [J]. Mol Ecol Resour, 2008, 8(1): 103-6.

[5] GOUY A, ZIEGER M. STRAF-A convenient online tool for STR data evaluation in forensic genetics [J]. Forensic Sci Int Genet, 2017, 30: 148-51.

[6] DELEST A, GODFRIN D, CHANTREL Y, et al. Sequenced-based French population data from 169 unrelated individuals with Verogen's ForenSeq DNA signature prep kit [J]. Forensic Sci Int Genet, 2020, 47: 102304.

[7] GETTINGS K B, APONTE R A, VALLONE P M, et al. STR allele sequence variation: Current knowledge and future issues [J]. Forensic Sci Int Genet, 2015, 18: 118-30.

[8] NOVROSKI N M M, KING J L, CHURCHILL J D, et al. Characterization of genetic sequence variation of 58 STR loci in four major population groups [J]. Forensic Sci Int Genet, 2016, 25: 214-26.

[9] 张驰,郭立亮,周轲等.运用法医学二代测序技术侦破19年久冷案[J].刑事技术,2022,47(01):100-106.

[10] 汤晓, 刘宗伟, 黄嘉伟等. 综合应用Y-STR和全同胞关系破获命案积案1例 [J]. 中国法医学杂志, 2021, 36(01): 114-5.

[11] 徐曲毅, 刘琦, 刘宏等. 利用地缘性犯罪Y-STR库认定11年双命积案嫌疑人 [J]. 中国法医学杂志, 2020, 35(03): 330-1.

[12] 刘宇轩, 李家富, 韦甜. 利用Y-STR家系排查破获多年前命案1起 [J]. 法医学杂志, 2020, 36(01): 144-6.