四川大学学报（医学版） 2021，52（4）：671－678
J Sichuan Univ（Med Sci） doi: 10.12182/20210760107

·新技术新方法·

一种基于主流单倍型的家系分类法以及基于贝叶斯理论的家系Y-STR容差规律研究*

臧正卿[1]，赵永红[1]，蹇 慧[2]，郝宏蕾[3]，苏艳佳[3]，梁伟波[2△]

1. 四川大学吴玉章学院/数学学院（成都 610065）；2. 四川大学华西基础医学与法医学院 法医物证学教研室（成都 610041）；
3. 浙江省公安物证鉴定中心 浙江省刑事科学技术应用研究重点实验室（杭州 310009）

【摘要】 目的 建立在大范围人群中区别不同男性家系的分类方法，研究中国汉族男性家系成员之间Y-STR基因座容差的分布规律，探索不同容差的个体对之间间隔不同减数分裂次数的概率分布情况。**方法** 收集12个中国汉族男性家系269名个体外周血样本与45名无关人员外周血样本，采用Yfiler Plus™与ZGWZ FSY或Yfiler Platinum试剂盒，获得314个Y-STR单倍型；以重复次数为3次及以上的Y-STR单倍型为主流单倍型，选择众数最大的主流单倍型作为第一类数据中心，按Y-STR分型容差在5个基因座且6个步长以内的标准进行聚类合并，再以剩余数据中众数最大的主流单倍型作为第二类中心，依次聚类；将家系成员和无关个体分别进行两两比对，统计家系成员之间和无关个体之间的容差分布情况，进一步计算各基因座平均容差率，利用贝叶斯公式计算不同容差条件下间隔不同减数分裂次数的概率分布情况。**结果** 269名个体被划分为12个群组，组内个体与12个已知家系成员数据的对应率为100%，45名无关个体呈散点分布；家系成员之间的容差基因座数目分布在0～7个基因座与0～7个步长以内，无关个体之间的差异则至少在11个基因座和15个步长及以上；各家系内部一步容差和两步容差数目最多的基因座均各不相同，具有家系特异性；各基因座最小突变次数、平均容差率均与突变率显著相关；0容差的两个体有19.7%的概率间隔1次减数分裂，有71.2%的概率间隔6次以内；3个一步容差的两个体有65.2%的概率间隔减数分裂次数为10次以上。**结论** 以主流单倍型为聚类中心的聚类方法可以对大规模男性家系样本进行快速有效的区分，以及从中获得的不同容差条件下间隔不同减数分裂次数的概率分布情况，可为今后利用Y-STR数据库在家系调查、数据分析与实战应用中提供研究思路、筛选工具和重要参考依据。

【关键词】 法医物证学 Y-STR 容差 主流单倍型 聚类分析法 男性家系

Method of Identifying Male Lineages Based on Main Haplotype and Analysis of the Distribution of Y-STR Haplotype Mismatch Based on the Bayesian Theory

ZANG Zheng-qing[1], *ZHAO Yong-hong*[1], *JIAN Hui*[2], *HAO Hong-lei*[3], *SU Yan-jia*[3], *LIANG Wei-bo*[2△]. 1. *Wu Yuzhang Honors College and College of Mathematics*, *Sichuan University*, Chengdu 610065, China; 2. *Department of Forensic Genetics*, *West China School of Basic Medical Sciences and Forensic Medicine*, *Sichuan University*, Chengdu 610041, China; 3. *Zhejiang Provincial Key laboratory of Application Research in Forensic Science and Technology, Institute of Forensic Genetics, Zhejiang Provincial Public Security Bureau*, Hangzhou 310009, China

△ Corresponding author, E-mail: liangweibo@scu.edu.cn

【Abstract】 Objective To establish a classification method to identify different male lineages in a large population, to study the distribution patterns of Y-STR loci mismatches among Han Chinese male lineage members and to explore the mismatch probability distribution among the members with different meiosis intervals in the family. **Methods** Peripheral blood samples of 269 male individuals from 12 lineages in Han Chinese population and 45 unrelated male individuals were collected. Then, Yfiler Plus™ and ZGWZ FSY or Yfiler Platinum amplification kits were used, obtaining 314 Y-STR haplotypes. The Y-STR haplotype with 3 or more repetitions were selected as the main haplotype, in which the largest number was selected as the first data center. According to the standard of Y-STR genotype, those with mismatches within five loci and six steps were clustered and merged. Then, the main haplotype of the largest number in the remaining data was taken as the second data center, and cluster analysis is carried out in turn until there is no main haplotype remained. Pair comparison was conducted between lineage members and unrelated individuals, and the mismatch distribution among lineage members and unrelated individuals was calculated respectively. The average mismatch rate of each locus was subsequently calculated, as well as the mismatch probability distribution among members with different meiosis intervals within the lineage. **Results** 269 out of the 314 individuals were divided into 12 groups by cluster analysis method, accomplishing 100% accuracy between the cluster groups thus identified and the 12 known lineages. The remaining 45 unrelated individuals were scattered. The mismatch loci was within 0-7 loci and 0-7 steps among lineage members and the mismatch between unrelated individuals was at least 11 loci and 15 steps. The mismatch loci with the largest number of one-step and two-step mismatch were different in each lineage and had features that were

* 国家自然科学基金（No. 81971799）资助

△ 通信作者，E-mail: liangweibo@scu.edu.cn

specific to each lineage. The minimum mutation count and average mismatch rate of each locus were significantly correlated with the mutation rate. Two individuals with no mismatch had a 19.7% probability of 1 meiosis interval and a 71.2% probability of less than 6 meiosis interval. Two individuals with 3 loci mismatches had a 65.2% probability of more than 10 meiosis intervals. **Conclusion** The cluster analysis method based on main haplotypes provided in this paper can quickly and effectively differentiate large male lineage samples. The clustering method and the mismatch probability distribution of different meiosis intervals obtained thus can provide new ideas for research and screening instruments, and important reference for lineage investigation, data analysis and practical application of Y-STR database in the future.

【Key words】 Forensic genetics Y-STR Mismatch Main haplotype Cluster analysis Male lineage

Y染色体遗传标记具有男性特有、父系遗传等特点[1]。同一父系成员共享同一种单倍型,因此Y染色体遗传标记被广泛应用于法庭科学中[2-3]。Y染色体数据库已经在全国范围内建立并逐渐应用于实战[4-5],一些重大案件经常需要开展大范围的调查与采样,分析Y染色体遗传标记来搜索嫌疑人[6-8]。但多数情况下,比中数据并不一定是嫌疑人本人,定位嫌疑人可能来源的家系,对案件的调查可提供重要意义[9]。Y-STR在相对稳定的情况下有一定的突变率(4.0×10^{-3})[10-11],突变会在传代过程中随机发生,造成同一家系男性个体之间的Y-STR单倍型并不完全一致,即家系成员之间Y-STR分型容差现象较为普遍,这为区分近几代内的不同男性家系个体提供可能性,但这也会使得在进行Y数据库比对时获得的比对结果更加复杂和难以解释[12-13]。实际工作中,在面对大量的Y-STR单倍型数据时,如何快速有效地进行家系分类?此外,研究真实情况下容差在家系内外的分布规律,进而探索不同容差的个体对之间可能的间隔减数分裂次数,均是现阶段亟待解决的问题。本研究收集12个已知家系来源的269名中国汉族男性个体与45名无关样本个体血样,针对目前法庭科学实验室常用的35个Y-STR基因座,采用Yfiler Plus™与ZGWZ FSY,或Yfiler Platinum试剂盒进行Y-STR分型,首先以主流单倍型[14]为聚类中心对314名个体的Y-STR单倍型进行聚类,建立能够快速区别不同家系的新方法;其次,研究家系成员和无关个体之间的容差分布规律,研究突变率、家系样本组成情况等因素对容差分布造成的影响,探索家系成员之间不同容差数与间隔减数分裂次数的概率分布,为今后Y-STR数据库在家系调查、数据分析与实战应用中提供研究思路、筛选工具和科学依据。

1 材料与方法

1.1 样本

本研究经四川大学医学伦理委员会批准(KS2019042)。经调查确认来源于12个家系的269名中国汉族人群男性血样本,为本实验室日常办案与建库积累,样本组成情况见表1。其中有5个家系(家系5、11、13、14、15)共85名个体的系谱图明确,成员之间的间隔减数分裂次数分布在1~12次之间,具体情况见表2。另外收集45名无关人员样本,经调查确认与269名家系成员均无血缘关系。

表 1 12个已知中国汉族男性家系成员样本组成表

Table 1 The sample distributions of 12 male lineages of Chinese Han population

Lineage	Samplesize	Individual pairs	Meiosis interval
1	29	406	Pedigree unknown
2	51	1 275	Pedigree unknown
3	14	91	Pedigree unknown
5	31	465	1-12
6	31	465	Pedigree unknown
7	30	435	Pedigree unknown
8	16	120	Pedigree unknown
9	13	78	Pedigree unknown
11	20	190	1-11
13	6	15	2-3
14	14	91	1-8
15	14	91	1-7
Accumulation	269	3 722	-

表 2 已知系谱图的5个家系成员之间不同间隔减数分裂次数分布表

Table 2 The distribution of individual pairs between different meiosis intervals among 5 lineages with known pedigree

Meiosis interval	Lineage 5	Lineage 11	Lineage 13	Lineage 14	Lineage 15	Accumulation
1	1	3	0	2	5	11
2	13	11	10	9	6	49
3	15	5	5	4	8	37
4	15	10	0	0	9	34
5	24	1	0	20	19	64
6	44	32	0	41	31	148
7	14	47	0	14	13	88
8	28	43	0	1	0	72
9	9	25	0	0	0	34
10	85	10	0	0	0	95
11	175	3	0	0	0	178
12	42	0	0	0	0	42

1.2 方法

1.2.1 Y-STR分型检测　所有血样采用直扩法进行Yfiler Plus™(美国Thermo Fisher Scientific公司)和ZGWZ FSY(浙江省公安物证鉴定中心)试剂盒, 或者Yfiler Platinum(美国Thermo Fisher Scientific公司)试剂盒的扩增检验, 10 μL扩增体系, 各试剂成分配比与扩增循环参数按照各试剂盒说明书, 扩增反应在9700型PCR扩增仪(美国Thermo Fisher Scientific公司)上进行, 扩增产物使用3500XL型基因分析仪(美国Thermo Fisher Scientific公司)检测, 采用GeneMapper IDX v1.5软件分析获得Y-STR分型数据。上述试剂盒共同的35个Y-STR基因座包括DYS389I、DYS448、DYS389II、DYS19、DYS391、DYS481、DYS549、DYS533、DYS438、DYS437、DYS635、DYS390、DYS439、DYS392、DYS643、DYS393、DYS385ab、DYS456、Y-GATA-H4、DYS460、DYF387S1、DYS527、DYS444、DYS557、DYS447、DYS522、DYS576、DYS570、DYS458、DYS627、DYS518、DYS449, 其突变率取自7篇群体遗传学研究论文[10, 15-20]中获得的累计5 792名中国汉族男性的突变率均值。其中3个基因座(DYS518、DYS576和DYS627)为快速突变基因座, 其余32个为中低突变。

1.2.2 以主流单倍型为中心的家系聚类分析方法　对314名个体单倍型以主流单倍型为聚类中心进行聚类分析。记单倍型数据分别为$H_i, i=1,2,\cdots,n$, H_i和H_j的容差基因座数记为l_{ij}, H_i和H_j的容差步数记为s_{ij}, 记C_i表示经聚类分析得到的第i类。聚类步骤如下:

第一步: 寻找第一类的聚类中心, 本研究选取众数最大的主流单倍型H_{i_1}作为聚类中心, 即

$$\text{mode}(H_{i_1})=\max\{\text{mode}(H_1),\text{mode}(H_2),\cdots,\text{mode}(H_n)\},\quad \text{mode}(H_{i_1})>1。$$

如果出现$\text{mode}(H_{i_1})=\text{mode}(H_{i_2})=\cdots=\text{mode}(H_{i_k})$, 则按照$i_1<i_2<\cdots<i_k$的顺序, 将$H_{i_k}$作为第$k$个聚类中心。

第二步: 对主流单倍型定义距离向量:

$$\vec{d}(H_j,H_{i_1})=(l_{j_{i_1}},s_{j_{i_1}}),j=1,2,\cdots,n。$$

第三步: 根据本研究2.2.1部分的结果, 同一家系成员之间, 绝大部分(99%)容差分布在5个基因座与6个步长及以内, 因此本部分设定以下判别规则:

如果有$\{l_{j_{i_1}}\leqslant 5\}$且$\{s_{j_{i_1}}\leqslant 6\}$, 则$H_j\in C_1, j=1,2,\cdots,n$, 第一类聚类结束。

第四步: 剔除第一类C_1的个体后, 重复上述过程直到聚类结束。

首先使用本研究中的基于主流单倍型的家系分类法对样本进行分析后, 将获得的聚类分析结果利用Network10软件[21]进行结果展示。将获得的聚类分析结果以及每个单倍型基因座信息、该单倍型的众数都录入Network10的后缀为".ych"的文件中; 然后, 用该ych文件进行Median Joining步骤生成out文件, 再用Network10中的Draw Network步骤生成网络图。

1.2.3 容差规律探索　使用直接计数法统计35个Y-STR基因座在12个家系内部的一步容差、两步容差的分布情况, 并探索其与各家系个体数之间的关系; 结合样本实际分型, 推断家系内基因座实际发生的最小突变次数, 以及与各基因座突变率进行相关性分析。

计算各基因座在各家系的容差率, 获得各基因座平均容差率, 并与各基因座突变率进行相关性分析。

计算不同容差数对应的间隔不同减数分裂次数的概率。将5个已知系谱图的家系成员的单倍型进行两两比对, 使用直接计数法统计不同间隔减数分裂次数的个体对的家系组成情况, 以及不同间隔减数分裂次数的个体对的容差分布情况。

记间隔减数分裂次数为离散型随机变量X, X可能取值为1 ~ 12, 容差数Y也是离散型随机变量, 已知系谱图的家系数为m。令n_{ijk}表示第k个家系内间隔i次减数分裂且容差数为j的个体对数, n_{ik}表示第k个家系内间隔i次减数分裂的个体对总数, 记间隔i次减数分裂的条件下容差数为j的概率为$P(Y=j|X=i)$, 则

$$P(Y=j|X=i)=\frac{1}{m}\sum_{k=1}^{m}\frac{n_{ijk}}{n_{ik}},i=1,2,\cdots,12,j\in N。$$

由贝叶斯公式, 可估计在不同容差数下间隔不同减数分裂次数的概率:

$$P(X=i|Y=j)=\frac{P(X=i)P(Y=j|X=i)}{\sum_{i=1}^{12}P(X=i)P(Y=j|X=i)},\quad i=1,2,\cdots,12,j\in N。$$

假设各间隔减数分裂次数X的先验概率为等可能概型, 即

$$P(X=i)=\frac{1}{12}(i=1,2,\cdots,12)\ 。$$

则上述公式可简化为:

$$P(X=i|Y=j)=\frac{P(Y=j|X=i)}{\sum_{i=1}^{12}P(Y=j|X=i)}i=1,2,\cdots,12,j\in N\ 。$$

由此可估计不同容差数条件下的间隔不同减数分裂次数的概率分布。

2 结果

2.1 聚类结果

2.1.1 314名个体经统计获得的主流单倍型及众数分布　对来源于12个家系的269名中国汉族人群男性血样本与

45个无关样本进行统计分析，发现其中有21种主流单倍型，其众数分布从3～15次不等。

2.1.2 以主流单倍型作为聚类中心的个体聚类结果 314名个体单倍型经聚类分析，将获得的聚类分析结果利用Network10软件进行展示，结果如图1所示。从图1可见，314名个体中269名来自12个家系的个体被相应的聚类到12个群组，而45个无关个体形成45个离散的点。

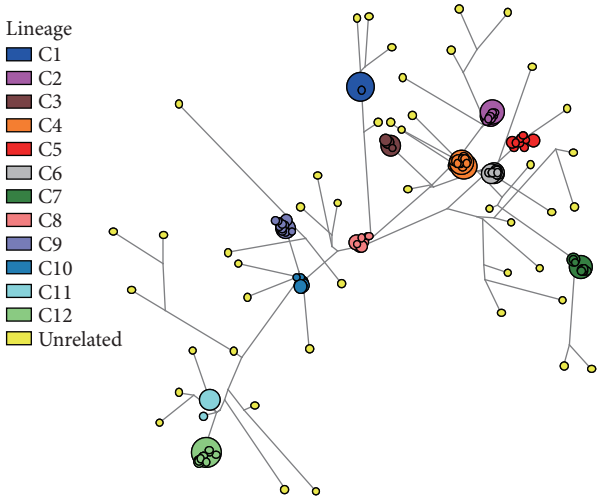

图 1 314名个体Y-STR分型聚类结果图

Fig 1 The results of cluster analysis from the Y-STR of 314 individuals

The 12 clusters with different colors and discrete points in the figure respectively represent the distribution of family members and the distribution of unrelated individuals after cluster analysis, and the size represents the number of people within a particular classification.

2.2 容差规律统计结果

2.2.1 个体之间Y-STR单倍型容差统计结果 同一家系成员之间的容差分布在0～7个基因座与0～7个步长以内，绝大部分(99%)容差在5个基因座与6个步长及以内，无关个体之间的差异则至少在11个基因座和15个步长及以上，两者分布差异明显，且没有交集(图2)。

2.2.2 12个家系内部各基因座容差分布情况 分析12个家系组成样本量及35个Y-STR基因座累计容差分布和两步容差情况。首先对12个家系内部269名人员在35个基因座上分别进行两两比较，共130 270次，观察到7 947次容差，发生容差的概率约为6.1%，其中一步容差7 665次(96.45%)，两步容差281次(3.54%)，三步容差1次(0.01%)。除家系13没有发现容差，其余家系均有发现。仅DYS391和DYS438两个基因座没有观察到容差，其余32个基因座均有发生容差。4个家系(家系5、6、14、15)出现容差数最多的基因座为快突变基因座(DYS518、DYS576和DYS627)，其余各家系容差数最多的基因座均各不相同。

如表3所示，12个家系中有8个家系在8个基因座上出现两步容差，4个家系(家系3、11、13、14)未发现两步容差。不同家系出现两步容差的基因座各不相同，并且与各基因座突变率没有显著相关性。唯一发生三步容差的是家系6的DYS627基因座，该家系内有29个两步容差和1个三步容差及183个一步容差。

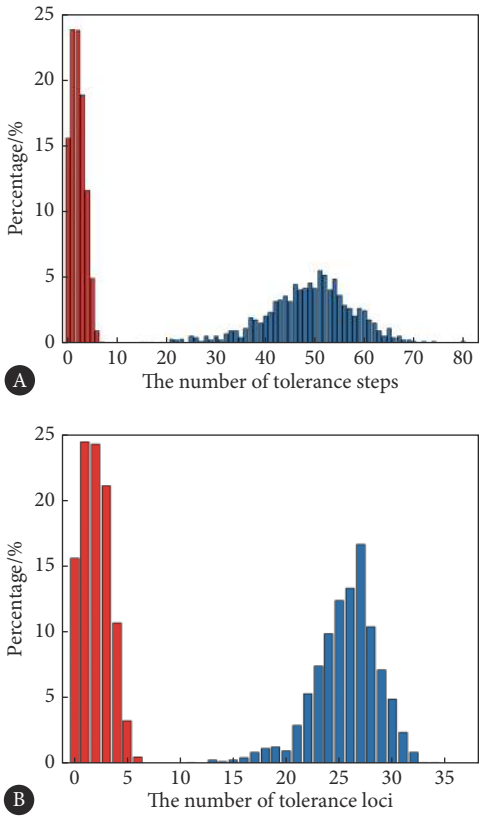

图 2 同一家系与无关个体之间的容差基因座数与容差步数分布图

Fig 2 The distribution of mismatch loci and mismatch steps between the same male lineage and independent individuals

A is the distribution result of tolerance locus number between the same family and unrelated individuals; B is the distribution result of tolerance steps between the same family and unrelated individuals. Red represents family members and blue represents unrelated individuals. There is no crossover between family members and unrelated individuals in the two figures.

2.2.3 35个Y-STR基因座在12个家系中的最小突变次数 容差本质上是由突变导致，根据容差分布情况和家系样本数，可以初步推测各基因座实际发生的最小突变次数。以家系5为例，该家系包含31个样本，在DYS389Ⅱ、DYS19、DYS392、DYS385ab、DYS456、Y-GATA-H4这6个基因座各有30个容差，经过对个体分型以及系谱图的仔细研究，发现该家系内部，这6个基因座，31个样本中各仅1个样本携带突变分型，因此两两比对各造成30个容差，即该家系内这6个基因座各自发生突变的最小次数均为1次；DYS570有58个容差，有两个不同分支的样本携带同一种突变分型，推测该基因座在该家系中可能最少发生

表 3　12个家系内两步容差分布情况

Table 3　The distribution of two-step mismatches in 12 lineages

Lineage	Sample size	Individual pairs	DYS549	DYS389 Ⅱ	DYS627	DYS458	DYS449	DYS390	DYS635	DYS385ab	Accumulation
1	29	406	0	0	0	0	17	0	0	0	17
2	51	1 275	0	0	0	125	3	2	0	0	130
5	31	465	0	0	20	0	0	0	0	0	20
6	31	465	0	4	29	0	0	0	56	0	89
7	30	435	0	0	0	3	0	0	0	0	3
8	16	120	4	0	0	0	0	1	0	0	5
9	13	78	0	0	0	0	0	0	0	15	15
15	14	91	0	0	0	0	2	0	0	0	2
Accumulation			4	4	49	128	22	3	56	15	281

了两次突变。而DYS627基因座, 有3个不同分支的4个个体分型为22, 4个不同分支的5个个体分型为24, 其余22个个体分型为23, 因此两两比对时, 出现20个两步容差和198个一步容差, 推测该基因座在该家系内实际发生的最小突变次数为7次。

使用该方法, 通过分析各家系样本数和各基因座发生一步容差、两步容差的数目, 结合系谱图以及家系内样本实际分型, 推断家系内部各基因座实际发生的最小突变次数。结果显示, 突变发生最多的基因座为快速突变DYS627基因座(14次), 各基因座最小突变次数与突变率之间的皮尔逊相关性系数为0.861, $P<0.001$, 两者呈显著相关。

2.2.4　35个Y-STR基因座的平均容差率　用各基因座家系内累计容差数除以家系内个体对数, 获得各基因座在每个家系的容差率, 再求12个家系各基因座容差率的均值, 获得各基因座平均容差率, 结果见表4。12个家系平均容差率最高的基因座是快速突变基因座DYS627。各基因座平均容差率与突变率之间的皮尔逊相关性系数为0.745, $P<0.001$, 两者呈显著相关。

2.2.5　不同容差数对应的不同间隔减数分裂次数的概率　通过对5个系谱图已知的家系(家系5、11、13、14、15)进行分析, 统计不同容差数中不同间隔减数分裂次数的个体对数的分布情况, 结果见表5。间隔1次减数分裂的个体对最少(11对), 且均是0容差; 间隔11次减数分裂的样本对数最多(178对), 其中59对是3个一步容差; 其次是间隔6次减数分裂的样本对(149对), 其中85对为0容差。

通过分析间隔不同减数分裂次数的个体对的家系组成情况以及各家系内该间隔减数分裂次数的个体对总数, 先计算间隔i次减数分裂的条件下容差数为j的概率, 即$P(Y=j|X=i)$。根据方法1.2.3中贝叶斯公式计算获得不同容差数条件下间隔不同减数分裂次数的概率, 结果见表6。如果两个体35个Y-STR为0容差, 间隔12次减数分裂的概率为0; 1个一步容差和2个一步容差的两个体分别有13.5%和15.3%的概率间隔4次减数分裂, 3个一步容差、4个一步容差、5个一步容差的两个体分别有26.5%、28.8%和24.9%的概率间隔12次减数分裂。将1 ~ 3、4 ~ 6、7 ~ 9、10 ~ 12次间隔减数分裂次数的概率分别进行合并, 结果如图3所示。如果两个体35个Y-STR为0容差, 两者有

表 4　35个Y-STR基因座平均容差率

Table 4　Average mismatch rate of 35 Y-STR loci

Locus	Average mismatch rate ($\times10^{-2}$)	Locus	Average mismatch rate ($\times10^{-2}$)
DYS438	0	DYS439	3.188 182
DYS437	0.574 713	DYS390	2.551 523
DYS448	0.537 634	DYS456	1.112 347
DYS392	0.537 634	DYS481	7.107 723
DYS643	2.197 802	DYS389 Ⅱ	4.846 358
DYS393	1.190 476	DYS385ab	6.822 954
DYS19	2.154 014	DYS385ab-2	15.805 580
DYS389 Ⅰ	1.505 376	DYS570	7.441 598
DYS533	0.326 797	DYS458	11.714 780
DYS549	7.745 374	DYF387S1	3.458 802
Y-GATA-H4	1.075 269	DYF387S1-2	7.249 431
DYS522	6.240 790	DYS527	3.616 271
DYS391	0	DYS527-2	1.903 858
DYS635	4.572 354	DYS449	14.568 320
DYS460	4.120 879	DYS518	11.037 830
DYS557	1.370 968	DYS576	13.184 630
DYS444	2.695 853	DYS627	16.237 620
DYS447	6.820 902		

表 5 不同容差数中不同间隔减数分裂次数的个体对分布情况

Table 5 The distribution of individual pairs with different meiosis intervals in different mismatch loci

Meiosis interval	Individual pairs	0 mismatched locus	1 mismatched locus	2 mismatched loci	3 mismatched loci	4 mismatched loci	5 mismatched loci	6 mismatched loci	7 mismatched loci
1	11	11							
2	49	31	12	4	1			1	
3	37	17	14	5				1	
4	34	14	12	7					1
5	64	22	31	9	2				
6	149	85	40	16	8				
7	88	18	21	10	16	11	6	5	1
8	72	12	25	16	12	5	2		
9	34	6	7	6	10	2	3		
10	95	8	22	14	30	13	6	2	
11	178	42	20	21	59	24	10	2	
12	42		10	2	21	6	3		
Accumulation	853	266	214	110	159	61	30	11	2

表 6 不同容差数的个体对之间间隔不同减数分裂次数的概率

Table 6 The probability of different meiosis intervals between individual pairs with different mismatch loci

Meiosis interval	0 mismatched locus	1 mismatched locus	2 mismatched loci	3 mismatched loci	4 mismatched loci	5 mismatched loci	6 mismatched loci	7 mismatched loci
1	19.69%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
2	12.25%	9.79%	5.29%	0.96%	0.00%	0.00%	21.13%	0.00%
3	11.48%	10.21%	8.72%	0.00%	0.00%	0.00%	18.32%	0.00%
4	7.73%	13.48%	15.34%	0.00%	0.00%	0.00%	0.00%	80.69%
5	9.85%	12.98%	8.21%	1.40%	0.00%	0.00%	0.00%	0.00%
6	10.16%	10.56%	9.27%	3.35%	0.00%	0.00%	0.00%	0.00%
7	5.89%	11.60%	9.63%	6.58%	11.79%	11.13%	36.54%	19.31%
8	8.24%	8.48%	12.64%	6.25%	7.81%	5.41%	0.00%	0.00%
9	2.36%	7.61%	13.89%	16.26%	22.39%	20.92%	0.00%	0.00%
10	3.53%	4.66%	9.19%	21.05%	15.41%	27.69%	16.16%	0.00%
11	8.81%	2.06%	4.36%	17.62%	13.82%	9.96%	7.85%	0.00%
12	0.00%	8.58%	3.46%	26.51%	28.79%	24.90%	0.00%	0.00%
Accumulation	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

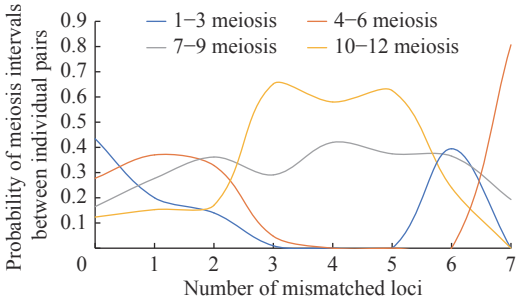

图 3 不同容差数的个体对之间间隔不同减数分裂次数的概率

Fig 3 The probability of different meiosis intervals between individual pairs with different mismatch loci

43.4%的概率间隔减数分裂次数在3次以内，有71.2%的概率间隔6次以内，而1个一步容差的两个体有37.0%的概率间隔减数分裂次数在4～6次；2个一步容差的两个体有36.2%的概率间隔减数分裂次数在7～9次；3个一步容差的两个体有65.2%的概率间隔减数分裂次数为10次以上。

3 讨论

目前Y-STR家系排查方法已被广泛应用于法庭科学领域，通过检测获得物证Y-STR分型，录入数据库进行比对，搜索物证检材可疑的家系来源[22]。数据库越大，越有

可能比对到越多的潜在家系。此时，如何快速有效对大量单倍型数据进行家系聚类分析，以及如何判断物证单倍型与比对到的单倍型之间的亲缘关系远近，是现阶段亟待解决的问题。

本研究采集了12个家系共269名家系人员样本，以及45名无关个体样本。对于家系个体分类研究，若采用经典的k均值聚类法[23]，由于k均值聚类法需预先设定聚类个数，在未知聚类个数的时候无法很好的将众多样本较好的分类，并且无法将无关个体样本区分开。本研究创新性地以主流单倍型为聚类中心，按照容差基因座数≤5且容差步数≤6判为同一聚类的标准进行聚类。获得的结果与原侦查获得的系谱图进行比较，对应率为100%，并且45名无关个体没有被错误的聚类到12个家系中，呈现散点分布的状态。提示本研究建立的以主流单倍型为数据中心的聚类分析法能够用于大规模男性家系个体分类，为大样本量家系数据的分类和整理提供帮助。陈炜结合Y-SNP和Y-STR进行父系溯源，利用Y-SNP可以实现大尺度上的家族关系的划定[24]。但是现有的Y库大部分是基于Y-STR分型数据建立起来的，本研究利用Y-STR进行家系分类，具有更高的实践价值。

但是，本研究给出的聚类分析方法，在出现某单倍型A跟其他两个单倍型B、C的容差基因座数和容差步数均相同，且均满足容差基因座数≤5且容差步数≤6的情况时，会将A判给主流单倍型众数更大的那一个家系，可能会导致错判。在以后的研究中需要考虑各基因座等位基因分布频率和突变情况，利用随机模拟的方法生成数据，比较B、C家系包含A个体的概率大小进行判别。此外，对于两个主流单倍型极其相似的家系，调查无法追溯其亲缘关系的，使用本研究建立的家系聚类分析方法无法将其区分开来，可能会出现错判的情况。可以考虑添加更多基因座或其他新的距离判别（如马氏距离[23]、相似度计算[23]等）指标进行进一步区分，亦或是该两个家系之间确实存在一定的亲缘关系，有待进一步的深入研究。此外，本研究收集的家系样本主要分布在5～7代，且针对35个常见的Y-STR基因座。如果家系代数不同，检测的Y-STR基因座与基因座数不同，其容差基因座数与步数会相应发生改变[25]，在选取聚类中心和聚类标准时应做相应调整，以免发生错判。

张广峰等[26]研究了在超过十代的家系中，男性个体两两之间Y-STR单倍型的差异幅度，为我们的工作提供了指导。通过对家系内部容差分布规律进行研究，发现家系内容差分布具有家系特异性，与家系采样个体的组成紧密相关。家系5、6、15容差数最多的基因座均在DYS627，其余家系容差数最多的基因座均各不相同。本研究中使用的3个快突变基因座是DYS518、DYS576和DYS627，仅在家系5、6、14、15中出现容差数最多。

同一家系成员之间的容差分布在0～7个基因座与0～7个步长以内，无关个体之间的差异则至少在11个基因座和15个步长及以上，该研究结果与吴微微等[12]的研究结果具有一致性。相较于一步容差，两步容差和三步容差发生次数较少，其家系特异性现象更为突出。12个家系中有8个家系在9个Y-STR基因座发现了两步容差，但是发生两步容差的基因座各不相同，并且与基因座突变率没有显著相关性。仅家系5和6在快突变基因座DYS627发现了两步容差，其余均发生在中低突变基因座。此外，家系5、6和7具有相似样本量（31、31和30），但是两步容差分布具有明显差异（分别为20、89和3）。唯一发现三步容差的是家系6（样本量为31）的快突变DYS627基因座，该家系内该基因座除了1个三步容差外，还有29个两步容差和183个一步容差。经过对分型仔细分析，发现该家系内有1个个体分型为21，7个个体分型为22，22个个体分型为23，1个个体分型为24，因此两两比对造成29个两步容差、1个三步容差以及183个一步容差。由于家系6系谱图未知，因此仅能推测DYS627在家系的现有样本中最少发生了3次突变，无法获得其详细突变过程。因此后续研究中收集系谱图完整的家系样本具有重要意义。

由于本质上容差的发生是由于家系内部传代过程中发生了突变，而因条件所限家系采样具有非随机性，且各家系采样量也不相同，可能造成各家系容差分布不同，样本量越小越容易出现偏差，因此还需对包含更多样本的家系进行综合分析寻找规律。本研究通过对家系综合分析获得平均容差率，研究发现该平均容差率与基因座突变率具有显著相关性；通过结合家系样本量和容差数，推测各基因座的最小突变次数，结果显示也和突变率具有显著相关性，揭示了容差、样本量与突变三者之间的关系。但是由于12个家系仅5个家系系谱图已知，能够准确推断突变次数，其余家系只能推测得到最小突变次数。这些结果说明，在探索家系内容差规律时，家系数目越多，家系内采样量越多，越有利于观察到真实分布情况，同时完善的家系系谱图也具有重要价值。

目前，对Y-STR基因座突变率的相关研究较多，但很少有文献利用家系内外Y-STR容差分布规律进行亲缘关系判断的研究报导。李为哲等[27]用Y-STR检测结果进行network分析并构建网络图，探索群体内个体间亲缘关。本研究通过对5个系谱图已知的家系进行分析，获得了不同容差数两个体间对应的间隔减数分裂次数的概率分布

情况，能够为实际检案提供更加直接的参考依据。家系内采样原因造成的容差家系特异性会给计算结果造成较大偏差，因此在计算过程中考虑每个家系的样本组成情况尤为重要。由于容差6个和7个基因座的情况在本研究涉及的12个家系269例样本中观察到的次数较少，概率计算结果可能存在较大误差，后续实验需要进一步扩大样本量以获得更加准确的结果。本研究获得的不同容差数两个体间对应的间隔减数分裂次数的概率，可以为今后利用在Y-STR数据库比对结果时提供参考依据，开展相应范围的侦查工作。

综上所述，本研究建立的以主流单倍型为数据中心的聚类分析法，在大范围男性家系调查中可以有效区别不同家系；获得的家系内Y-STR容差分布规律，可为今后利用Y-STR数据库在家系调查、数据分析与实战应用中提供研究思路、筛选工具和评估依据。同时本研究也指出，更多的家系数、更多的家系样本、更完善的系谱图，对于Y-STR容差规律分析具有重要意义。

* * *

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] KAYSER M. Forensic use of Y-chromosome DNA: A general overview. Hum Genet, 2017, 136(5): 621–635.

[2] RALF A, LUBACH D, KOUSOURI N, *et al*. Identification and characterization of novel rapidly mutating Y-chromosomal short tandem repeat markers. Hum Mutat, 2020, 41(9): 1680–1696.

[3] ROEWER L, ANDERSEN M M, BALLANTYNE J, *et al*. DNA commission of the International Society of Forensic Genetics (ISFG): Recommendations on the interpretation of Y-STR results in forensic analysis. Forensic Sci Int Genet, 2020, 48: 102308[2021-02-01]. https://doi.org/10.1016/j.fsigen.2020.102308.

[4] 崔杨程，陈雪云，李维丽，等. Y-STR单倍型分析对男性家系数据库建设的探讨. 中国法医学杂志, 2020, 35(1): 34–37.

[5] 张颖，张黎，吴妍. Y-STR数据库建设工作的实践与思考. 中国刑事警察, 2019(5): 62–64.

[6] 刘亚举，张俊涛，孙现锋. Y-STR家系排查及数据库建设问题. 中国法医学杂志, 2015, 30(2): 223–225.

[7] 苏志强. 家系排查在“Y-STR”DNA数据库建设中的作用. 中国法医学杂志, 2017, 32(S1): 24–25.

[8] 包若瑜，杜盼新，石美森，等. Y-STR数据库和遗传关系网络图在强奸案中的应用1例. 法医学杂志, 2020, 36(3): 420–422.

[9] LIU H, LI X Y, MULERO J, *et al*. A convenient guideline to determine if two Y-STR profiles are from the same lineage. Electrophoresis, 2016, 37(12): 1659–1668.

[10] 吴微微，王怀锋，郝宏蕾，等. 中国汉族人群46个Y-STR基因座多态性与突变调查. 中国法医学杂志, 2015, 30(3): 256–259.

[11] CLAERHOUT S, VANDENBOSCH M, NIVELLE K, *et al*. Determining Y-STR mutation rates in deep-routing genealogies: Identification of haplogroup differences. Forensic Sci Int Genet, 2018, 34: 1–10.

[12] 吴微微，张晓霞，金雷，等. 中国汉族男性家系Y-STR基因座容差情况调查. 中国法医学杂志, 2020, 35(4): 390–393.

[13] ANDERSEN M M, BALDING D J. How convincing is a matching Y-chromosome profile? PLoS Genetics, 2017, 13(11): e1007028[2021-02-06]. https://doi.org/10.1371/journal.pgen.1007028.

[14] 吴微微，任文彦，郝宏蕾，等. Y-STR单倍型信息指导数据库采样分析. 刑事技术, 2013(1): 3–5.

[15] FU J, CHENG J, WEI C, *et al*. Assessing 23 Y-STR loci mutation rates in Chinese Han father–son pairs from southwestern China. Molecular Biology Reports, 2020, 47(10): 7755–7760.

[16] LIN H, YE Q, TANG P, *et al*. Analyzing genetic polymorphism and mutation of 44 Y-STRs in a Chinese Han population of Southern China. Legal Med(Tokyo), 2020, 42: 101643[2021-02-11]. https://doi.org/10.1016/j.legalmed.2019.101643.

[17] 朱传红，周翔，邵毅，等. 湖北地区汉族人群24个Y-STR基因座遗传多态性研究. 刑事技术, 2014(1): 16–22.

[18] 吴微微，郝宏蕾，任文彦，等. 中国汉族人群17个Y-STR基因座突变情况分析. 中国法医学杂志, 2012, 27(6): 455–457.

[19] WANG Y, ZHANG Y J, ZHANG C C, *et al*. Genetic polymorphisms and mutation rates of 27 Y-chromosomal STRs in a Han population from Guangdong Province, Southern China. Forensic Sci Int Genet, 2016, 21: 5–9.

[20] 吴微微，苏艳佳，梅兴林，等. 30个Y-STR基因座在中国汉族人群中的多态性与突变. 法医学杂志, 2018, 34(4): 411–416.

[21] BANDELT H J, FORSTER P, RÖHL A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol, 1999, 16(1): 37–48.

[22] 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. 中国法医学杂志, 2021, 36(1): 114–115.

[23] 高惠璇. 应用多元统计分析. 北京: 北京大学出版社, 2005: 176–251.

[24] 陈炜. 联用Y-SNPs和Y-STRs构建有效的父系溯祖手段. 昆明: 昆明理工大学, 2020.

[25] 张广峰，高珊，畅晶晶，等. Y-STR单倍型在大家系中的差异研究. 刑事技术, 2018, 43(2): 138–143.

[26] 张广峰. Y-STR位点家系突变的研究. 北京: 公安部物证鉴定中心, 2017.

[27] 李为哲，李钊，党钦，等. 利用Y-STR单倍型推断样品间亲缘关系. 中国法医学杂志, 2020, 35(3): 305–308.

（2021－03－15收稿，2021－06－21修回）

编辑 余 琳